Investigating Impacts of Whole Genome Duplication on Immunogenetic Diversity and Parasite Load in Corydoradinae Catfish

Ellen Alicia Bell

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia, UK

School of Biological Sciences

September 2018

Dedication

To my Friends and Family

Past and Present

Thank You

“I’d rather be a rising ape than a falling angel”

Terry Pratchett, 2009

Abstract

Whole genome duplication (WGD) events have occurred repeatedly in the evolutionary history of plant and, less commonly, animal lineages, but their role as a facilitator of evolution is still not fully understood. Whole genome duplication events have been identified in the early history of vertebrates, teleosts and angiosperms and have been hypothetically linked to the large-scale diversification that has been described in these lineages. The Corydoradinae catfishes are a highly diverse sub-family of Neotropical catfishes with over 170 species described. A key feature of this subfamily is their history of whole genome duplication events. Previous studies have divided the Corydoradinae into nine distinct lineages and Restriction site Associated DNA (RAD) sequencing data identified that lineages 2 to 9 had undergone a WGD event 35 to 66 MYA and lineages 6 and 9 had undergone a second WGD event 20 to 30 MYA. Species belonging to the Corydoardinae coexist in sympatric and often mimetic mixed species communities with representatives of two or more or the nine lineages. This makes them a novel animal system for exploring the effects of WGD.

It is well understood that hosts derive benefits from carrying immune genes with high levels of diversity. Polymorphisms in immune genes mean that there is a greater probability of detecting a wider range of pathogenic antigens and there is an intrinsic advantage in expanding the pathogenic repertoire that an immune system can respond to. Toll-like receptors (TLRs) are a type of pathogen recognition receptor that function as part of the innate immune system. We have compared TLR2 in single individuals across the nine Corydoradinae lineages. Results show that lineages 2 and 7 had very high levels of TLR2 diversity and that all lineages, except for lineage 1 and lineage 6, had retained more then two haplotypes of this gene. When examining TLR diversity in two TLR genes (TLR1 and TLR2) in two coexisting populations of *Corydoras*, *C. maculifer* (a lineage 1, diploid) and *C. araguaiaensis* (a lineage 9, putative tetraploid) we found greater functional diversity in *C. araguaiaensis*. We also found that *C. araguaiaensis* had retained four copies of these TLRs and had not, as is common in polyploids, lost additional haplotypes of the duplicated genes. Conversely *C. maculifer* had a surprisingly low genetic diversity in TLRs, comparable to that found in endangered and/or bottlenecked populations of other taxa. When looking at a greater suit of immune genes across the *C. maculifer* genome this lack of diversity in immune genes held true. After assessing the parasite burden of populations of these two species we found that although the proportion of infected individuals in *C. maculifer* and *C. araguaiaensis* were similar the intensity of the infection was higher in *C. maculifer*. The increased immune gene diversity and reduced parasite intensity in the putative tetraploid *C. araguaiaensis* may be rare direct evidence of the adaptive advantage of whole genome duplication.

Table of Contents

Dedication ..... **i**

Abstract ..... **ii**

Table of Tables ..... **vi**

Table of Figures ..... **vii**

Acknowledgments ..... **x**

Chapter 1: General introduction **1**

1.1 Roles and consequences of polyploidy **2**

1.1.1 Mechanisms of polyploid establishment 3

1.1.2 Consequences of Polyploidy 6

1.2 Host parasite interactions and fish immunity **9**

1.2.1 Background of host parasite interactions 9

1.2.2 Fish immunity 9

1.2.3 Immune tissues and cellular elements of fish 10

1.2.4 Proteomic mechanisms of fish immunity 11

1.2.5 Polyploidy and Immunity 12

1.3 The Corydoradinae as a study system **15**

1.4 Aims and objectives **17**

Chapter 2: Characterisation of Toll-like Receptor 2 across the nine *Corydoras* Lineages **19**

2.1 Introduction **20**

2.1.1 Aims and objectives 21

2.2 Methods **23**

2.2.1 Sampling and DNA extraction 23

2.2.2 PCR amplification, library preparation and sequencing 23

2.2.3 Library preparation and sequencing 26

2.2.4 Data processing and analysis 30

2.3 Results **33**

2.3.1 Sequencing, Data Cleaning and Quality Control 33

2.3.2 Variant calling 33

2.3.3 Haplotype number 37

2.3.4 Variant distribution 40

2.3.5 Variant sharing between the nine lineages 40

2.4 Discussion **44**

2.4.1 Conclusion 45

Chapter 3: Toll-like Receptor variation within diploid and polyploid *Corydoras* catfishes 47

3.1 Introduction 48

3.1.1 Aims and objectives 50

3.2 Methods 52

3.2.1 Sampling and DNA extraction 52

3.2.2 PCR amplification and library preparation 52

3.2.3 Data processing and analysis 52

3.3 Results 55

3.3.1 Sequencing, Data Cleaning and Quality Control 55

3.3.2 Variant calling 58

3.3.3 Haplotype number quantification 61

3.3.4 Toll-Like Receptor Structure 67

3.3.5 Variant distribution and frequency 67

3.4 Discussion 70

3.4.1 Higher diversity among individuals and across the population of C. araguaiaensis 70

3.4.2 Haplotype retention 71

3.4.3 Structural variation and the distribution of SNPs across TLRs 72

3.4.4 Conclusion 74

Chapter 4: Parasite community comparisons between diploid and polyploid *Corydoras* catfish hosts 75

4.1 Introduction 76

4.1.1 Aims and objectives 77

4.2 Methods 79

4.2.1 Host sampling 79

4.2.2 Parasite extraction, identification and enumeration 79

4.2.3 Parasite DNA extraction, PCR amplification and sequencing 79

4.2.4 Data processing and analysis 82

4.3 Results 84

4.3.1 Parasite prevalence, intensity and abundance 84

4.3.2 Parasite community analysis 91

4.3.3 Immune gene association analysis 96

4.4 Discussion 100

4.4.1 Conclusion 103

Chapter 5: Characterisation of pathogen recognition receptors in *Corydoras maculifer* **105**
5.1 Introduction **106**
5.1.1 Aims and objectives 107
5.2 Methods **108**
5.2.1 Sampling and DNA extraction 108
5.2.3 Sequencing, Assembly and Scaffolding 108
5.2.4 Quality checks and Annotation 109
5.2.5 Immune gene mining and analysis 109
5.3 Results **111**
5.4 Discussion **117**
5.4.1. Conclusion 118
Chapter 6: Final Synthesis **119**
6.1 Synopsis **120**
6.1.1 Characterising TLR2 across the nine lineages 120
6.1.2 Variation of TLR1 and TLR2 across diploid and polyploid Corydoras populations 121
6.1.3 Parasite communities across diploid and polyploid Corydoras populations 122
6.1.4 Characterising pathogen recognition receptors in the Corydoras maculifer genome 122
6.2 Further work **123**
References **126**
Appendix ..... **xii**

Table of Tables

Table 2.1: A summary of lineage, sequence data available and the natural geographical distribution of the Corydoradinae species analysed within this chapter. 22

Table 2.2: Primers used to amplify TLR2 across Corydoras samples 25

Table 2.3: Species and primer specific PCR annealing temperatures 25

Table 2.4: Library adaptors and barcodes used to for NextSeq sequencing and sample identification 28

Table 2.5: Sequence read retrieval from the single ended sequencing run 34

Table 3.1: Read retrieval each analysis stage following the single ended sequencing run in C. maculifer and C. araguaiaensis. 56

Table 3.2: Averaged observed and expected heterozygosity metrics for TLR1 and TLR2 in C. maculifer (diploid) and C. araguaiaensis (putative tetraploid). 60

Table 3.3: Average Synonymous Non-Synonymous SNP counts per TLR in C. araguaiaensis, along with synonymous to non-synonymous SNP ratios (S:N). 60

Table 4.1: Universal primers used for PCR amplification of CO1 in nematode parasites (using IUPAC nucleotide ambiguity codes) 81

Table 5.1: Assembly, scaffolding and BUSCO summary statistics from the C. maculifer genome, demonstrating expected coverage and completeness. 113

Table 5.2: Summary mapping statistics for TLRs, NODs and NLRs in the C. maculifer genome along with predicted SNP and haplotype counts from QualitySNPng. 114

Table of Figures

Figure 2.1: The library preparation process including dual barcoding with barcode ligation and PCR annealed indices (not to scale). _____ *29*

Figure 2.2: Topology recovered from the phylogenetic analysis of protein alignments from TLR1 in C. maculifer and C. araguaiaensis, TLR2 across nine Corydoras lineages, and all known TLRs in D. rerio and I. punctatus.. _____ *35*

Figure 2.3: TLR2 SNP counts across the nine Corydoradinae lineages. Plot A shows average SNP counts across populations of C. maculfer (n= 17) and C. araguaiaensis (n=35) and Plot B shows total counts across single individuals of each species displayed. _____ *36*

Figure 2.4: Topology recovered from the phylogenetic analysis of TLR2 genes in nine Corydoras lineages, rooted to C. maculifer (lineage 1). _____ *38*

Figure 2.5: SNP read ratios within TLR2 in single individuals from lineages 2 to 8 and averaged across TLR1 and TLR2 in populations of lineage 1 (C. maculifer (n=17)) and lineage 9 (C. araguaiaensis (n=35)). _____ *39*

Figure 2.6: TLR2 domains inferred from SMART analysis and SNPs identified per species mapped according to amino acid position from representatives across the nine Corydoradinae lineages. _____ *41*

Figure 2.7: SNPs shared in at least two Corydoradinae lineages mapped according to their nucleotide position along TLR2. Lineages 2-8 were represented by single individuals, lineage 1 was represented by 17 individuals and lineage 9 was represented by 35 individuals _____ *42*

Figure 2.8: A) Number of SNPs shared across the nine Corydoradinae lineages in TLR2. Lineages 2-8 were represented by single individuals, lineage 1 was represented by 17 individuals and lineage 9 was represented by 35 individuals B) Number of SNPs shared across RAD sequence data from all nine lineages, each lineage represented by two individuals. _____ *43*

Figure 3.1: Topology recovered from the phylogenetic analysis of TLR1 (A) and TLR2 (B) genes in C. maculifer (tip label M) and C. araguaiaensis (tip label A) rooted with Ictalurus punctatus. _____ *57*

Figure 3.2: SNP counts across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) in two Toll-like receptor genes (TLR1 and TLR2) with SNPs divided by substitution type (i.e. synonymous and non-synonymous substitutions). _____ *59*

Figure 3.3: All TLR1 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP. _____ *62*

Figure 3.4: All Non-synonymous TLR1 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP. 63

Figure 3.5: All TLR2 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP. 64

Figure 3.6: All non-synonymous TLR2 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP. 65

Figure 3.7: Minimum inferred haplotype counts across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) at TLR1 and TLR2 as derived by QualitySNP and verified manually. 66

Figure 3.8: SNP read ratios averaged across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) at two TLR loci (TLR1 and TLR2). 66

Figure 3.9: A: Frequencies of alternative bases in TLR1 across populations of C. maculifer and C. araguaiaensis. Counted across individuals regardless of ploidy status. B: Alternative base frequencies in TLR1 across populations of C. maculifer and C. araguaiaensis weighted according to proportional presence based on read depth. 68

Figure 3.10: A: Frequencies of alternative bases in TLR2 across populations of C. maculifer and C. araguaiaensis. Counted across individuals regardless of ploidy status. B: Alternative base frequencies in TLR2 across populations of C. maculifer and C. araguaiaensis weighted according to proportional presence based on read depth. Assuming diploidy in C. maculifer and tetraploidy in C. araguaiaensis. 69

Figure 4.1: Prevalence (i.e. the proportion of infected hosts) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. 86

Figure 4.2: Intensity (i.e. the number of parasites per infected host) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. 87

Figure 4.3: Abundances (i.e. the number of parasites per host) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. 88

Figure 4.4: Standard length of host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), plotted against overall parasite count. 89

Figure 4.5: Predicted parasite abundances (number of parasites per host) per millimetre of host plotted according to host species and infected tissue type. 90

Figure 4.6: Intensity of parasites between two species of Corydoras catfishes, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), split according to tissue and parasite morphology. 92

Figure 4.7: Abundances of parasites between two species of Corydoras catfishes, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), split according to tissue and parasite morphology. 93

Figure 4.8: MDS visualisation of parasite communities and abundances across host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid). 94

Figure 4.9: Topology recovered from the phylogenetic analysis of nematode CO1 genes. 95

Figure 4.10: Non-synonymous SNP count in TLR1 and TLR2 of host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), plotted against overall parasite count. 97

Figure 4.11: Tree topology recovered from maximum likelihood analysis for TLR1 (A) and TLR2 (B) in C. araguaiaensis. 98

Figure 4.12: SNP association analysis for non-synonymous SNPs in TLR1 and TLR2 and parasite load in C. araguaiaensis. 99

Figure 5.1: Topology recovered from the phylogenetic analysis of known TLR, NOD and NLR associated genes across the C. maculifer genome (purple), available Corydoras species data, I. punctatus (channel catfish) and D. rerio (zebrafish). 115

Figure 5.2: Protein domain prediction for TLR, NOD and NLRs based on SMART analysis. 116

Acknowledgments

This thesis is the product of a fantastic journey, admittedly it has had its high and low points but by and large I have thoroughly enjoyed the experience. I have learnt a lot and I know I still have much to learn but I am excited to carry on the expedition. Science is a collaborative discipline, the final results presented here are the product of numerous networks, both professional and personal. I would not have been able to produce this thesis without the support and guidance of many people who I shall acknowledge here.

First of all, many thanks to my supervisors Martin (MIT) and David, you have both been very supportive over the last four years. You have both willingly shared your expertise and have never failed to make time for me when I needed it. Special thanks go to Martin because you also gave me the initial PhD opportunity. You gave me a shot at a PhD when so many others wouldn't and I was beginning to give up hope. I will never forget the moment I opened the PhD offer email from you, read it twice then burst into tears. For this I owe you much gratitude. I hope the next three years working together are even better then the last four have been.

In addition to my core supervisory team I have also benefited from the assistance and support of a number of collaborating labs. Many thanks are owed to Claudio Oliveira (CO) and his team for their help and expertise during sample collection and fieldwork in Brazil. The experience was a once in a lifetime opportunity and was hugely beneficial to both my personal and professional development. I would also like to thank Levi Yant and his lab group, especially Sarah, Sian, Christian, Patrick, Jordan and Pirita for all of their assistance with my library preparation and analysis in addition to all of their support with the genome sequencing and assembly. I feel like the success of the sequencing side of the project was largely due to the collaborative spirit of this lab and their willingness to take time out of their busy days to teach a novice the basics, thank you all. Many thanks are also owed to Jo Cable and her group of enthusiastic parasitologists. I found the parasitology side of this project very challenging but her groups knowledgeable nature, friendly good humour and palpable enthusiasm for my project put me at ease and gave me confidence in my work.

Many thanks are also owed to my family; my mum and dad (Rachel and Simon) and brother (Elliot), along with my grandmother (Olwyn) and grandfather (Rodney) who sadly passed away before I could submit this thesis, also my fury friends Henry and Nemo. You have all been unconditionally loving and supportive through this entire experience. You have tolerated the grumpy/sad phases and tried to ground me during the hyper/manic high points.

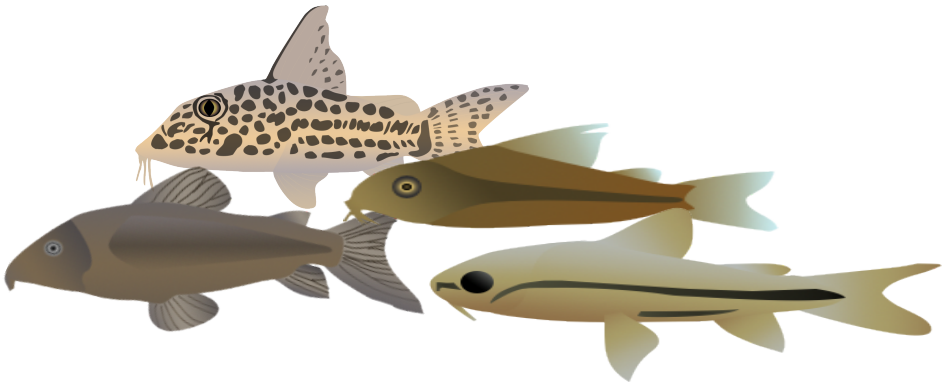
You have all been absolutely solid through this entire journey and provided a bedrock of support through the good and the bad times.

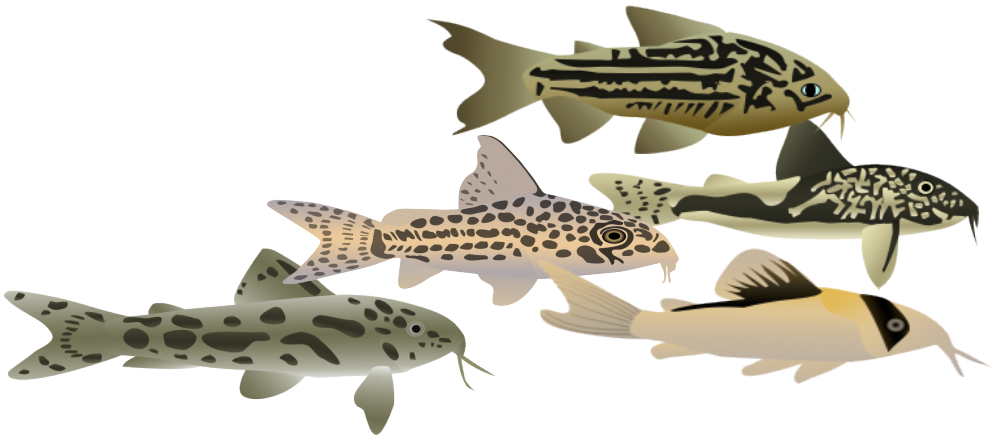
A number of people from UEA both past and present also deserve acknowledging. Sarah, you affectionately came up with the term “PhD sister” and that is exactly what you have been; you have taught me so much and been a guiding light when I was completely lost. You have also been an amazing friend and confidant. The Zeke office, Ents, Becky, Kat, Jen, James and Tom, you are the best office mates one could hope for, it’s an office full of laughter and mischief and some ridiculously random conversations. I would like to add a special mention for “little Ents”, you are a delightful mix of wisdom, excellent advise and wickedly naughty mischief, you make coming to work so much fun, don’t ever change. Also Becky, my favourite coffee buddy, I am so pleased I will get to see you finish your PhD as well, I will make sure I am on hand with the coffee whenever you need it. Jen and Collins, the dream team, thank you for listening to my endless rantings, your combined wit and wisdom are always a source of joy. Murray “Muzza”, my favourite grumpy Scott, you have been an exceptionally good listener throughout, you have also been a fantastic advisor whenever things got a little tricky and you have also never failed to bring a smile to my face. Dave and Laura, you have both been unshakably solid friends, providing unconditional support, I look forward to many more gameathons with you both.

A slightly unconventional acknowledgement should be made to my gym, BTF, whenever things have gotten really tough I have taken great pleasure in going to BTF and sweating it all out. The coaches Jon, Scott and Dean have never failed to push me physically even when I have been feeling mentally drained and been exceptionally grumpy. This has, at times been a huge sanity-restoring boost.

Thank you all for everything you have all done and please know that I am grateful.

Chapter 1:
General introduction





1.1 Roles and consequences of polyploidy

Whole genome duplication (WGD) events have been detected multiple times in the evolutionary history of animals and plants and have been implicated in major evolutionary transitions (Dehal and Boore, 2005; Peer, Maere and Meyer, 2009). For example, two rounds of WGD have occurred in the evolutionary history of vertebrates (the 2R hypothesis) and an additional fish specific whole genome duplication (FSGD) event in the teleost fish lineage (Dehal and Boore, 2005; Santini *et al.*, 2009). In plants as many as four WGD events have been detected in the evolutionary history of angiosperm lineages (Peer, Maere and Meyer, 2009). Historically the role played by WGD as an evolutionary mechanism has been debated. Some have argued that WGD events are of high importance as an evolutionary facilitator, leading to advances in species diversification and increasing biological complexity and novelty, while others maintain their significance is negligible (Stebbins 1940; Ohno, 1970 as cited in Furlong & Holland 2004; Peer et al. 2009). However in recent years, WGD/polyploidisation has been recognised more widely as having an important evolutionary role in plants and while this positive relationship is suspected in animals as well, mechanisms are less well understood (Mable, Alexandrou and Taylor, 2011).

Polyploidy has been defined as the presence of more than two sets of homologous chromosomes. Organisms are often identified as polyploid when chromosome numbers of closely related species follow a polyploidy associated series such as 2n= 16, 32, 64 etc (where 2n= somatic chromosome number) (Ramsey & Schemske, 1998), or less frequently when multivalents (an association of three or more chromosomes) are observed during meiosis (Otto and Whitton, 2000). Some authors have added to this definition, increasing its specificity to encompass chromosomal behaviours and gene expression. For example, Mable et al. (2011) defined polyploid species as “having twice the chromosome number of close relatives, sometimes retaining at least some pairing of multiple chromosome copies during meiosis and retaining evidence of duplicate gene expression distributed throughout the genome”. It is important to be aware of the various definitions of polyploidy and polyploid species, although within the context of this overview the classical definition has been adopted.

Historically, polyploidy has primarily been observed amongst plant species where it was estimated that as many as 70% of angiosperm species were polyploid (Masterson, 1994). However, incidences of polyploidy have also been observed in the animal kingdom, sporadically occurring in a number of vertebrate and invertebrate taxa including insects, fish, amphibians, reptiles and, possibly, a single mammal species (Otto and Whitton, 2000). Polyploid species are generally separated into either autopolyploids or allopolyploids. Traditionally, autopolyploids were described as arising within a single species whereas

allopolyploids were thought to be derived through inter-species hybridisation (Kihara & Ono 1926 as cited in Ramsey & Schemske 1998). Additional rules have been applied to these definitions, specifying that allopolyploids will form bivalents (associations of two chromosomes) during meiosis while autopolyploids form multivalents during meiosis(Otto and Whitton, 2000).

The timing of whole genome duplication (WGD) events is also of considerable interest to evolutionary biologists. It is possible to distinguish between taxonomic groups that have undergone ancient whole genome duplication (WGD) events (paleopolyploids) and groups where more recent WGD events have occurred. For example, it is now widely accepted that prior to their radiation (500-800MYA) vertebrates underwent two rounds of WGD; this is known as the 2R hypothesis (Dehal and Boore, 2005). It is also generally accepted that a fish specific whole genome duplication (FSGD) event occurred subsequent to the 2R event approximately 320-350MYA, when teleost fish began to diversify following their split from the *Holostei* (Santini et al., 2009). Teleosts encompass approximately 28, 872 species (99% of the total diversity of ray finned fishes) and make up the dominant radiation of vertebrates on the planet (Santini *et al.*, 2009). The FSGD event has been identified as a potentially causal factor influencing the wide radiation observed amongst teleost's (Santini *et al.*, 2009).

Following a whole genome duplication event, it is expected that most of the polyploid genome would slowly return to a diploid state, with a small proportion of duplicated genes retained (Petit *et al.*, 2004). This reduction process occurs because of a variety of mechanisms including; gene loss, silencing, fractionation (where a single gene copy is lost) along with genomic reorganisations such as chromosome fusion, fission, deletion or inversion (Soltis *et al.*, 2015; Robertson *et al.*, 2017). The mechanisms by which genes are lost and rediploidisation occurs tend to be species specific, at least among angiosperm species, and there are still substantial gaps our knowledge of the processes behind it (Soltis *et al.*, 2015).

1.1.1 Mechanisms of polyploid establishment

There are a number of mechanisms that may lead to the generation of polyploid organisms including: genomic doubling, gametic non-reduction and polyspermy. In addition to this there are also a number of factors that favour polyploid formation in organisms, these include: certain reproductive and chromosomal characteristics, gamete production, specific reproductive environments and the scope for hybridisation. Each of these factors will be discussed in greater detail here (Otto and Whitton, 2000; Mable, Alexandrou and Taylor, 2011).

The implications of genomic doubling and production of unreduced gametes in the context of polyploidy are similar. Genomic doubling relates to a failure of cell division following chromosome replication during mitosis (somatic polyploidy). The formation of unreduced gametes involves the failure of cells to undergo the second meiotic division during gamete production (Otto and Whitton, 2000). Both genomic doubling and unreduced gamete production have been reported in animals and plants. However, unreduced gametes appear to be more successful in generating polyploids when forming eggs or pollen, whereas unreduced sperm are thought to play a minor role in polyploidization, possibly because diploid sperm are less competitive compared to their haploid counterparts (Otto and Whitton, 2000). A further mechanism by which polyploidy may occur is polyspermy - where multiple sperm fuse with a single egg (Mable, Alexandrou and Taylor, 2011). Polyspermy has been observed in both plants and animals, although animals possess a range of physical and chemical mechanisms which prevent polyspermy from occurring (Mable et al. 2011; Otto & Whitton 2000).

Polyploids may also form, or be induced, through the retention of the second polar body following egg fertilization (Tiwary, Kirubagaran and Ray, 2005; Mable, Alexandrou and Taylor, 2011). It is possible to manipulate environmental variables to induce the formation of polyploids (Mable, Alexandrou and Taylor, 2011). In fish for example, triploidy has been induced by exposing fertilised eggs to temperature, hydrostatic pressure or chemical treatments; which may prevent extrusion of the second polar body or block the first mitotic division of the fertilised egg rendering the resulting progeny triploid (Tiwary et al. 2005, Mable et al., 2011). These processes have been used in a number of farmed fish (including Samlonids and Cyprinids) with an aim of inducing sterility and increasing growth and yield of fish farmed stocks (Zhou and Gui, 2017).

Once formed there are a number of obstacles newly arisen polyploids must navigate if they are to persist. Intrinsic issues arise in polyploid individuals due to imbalanced chromosome sets, incompatibilities between parental genomes, altered protein dosages and aneuploidy (partial change in the number of chromosomes in a chromosome set) (Mable, Alexandrou and Taylor, 2011). Disruptions to sex chromosome ratios may lead to sterility (or at least a reduction in fecundity) in polyploid individuals and may disrupt gender balances at the population level (Orr, 1990; Otto and Whitton, 2000). Based on difficulties surrounding the balancing of sex chromosomes in polyploid organisms, Otto & Whitton (2000) predicted that polyploidy would be most likely to occur in animal taxa with propensities for asexual or hermaphroditic reproduction, gender determination based on the presence or absence of a Y chromosome (as opposed to X chromosome to autosome ratio) and the presence of non-degenerate sex chromosomes with no dosage compensation.

The mechanisms described thus far explain how polyploid individuals may arise and the factors favouring their occurrence, however, newly arisen polyploids need to navigate a range of obstacles in order to establish themselves within a community. These obstacles include: niche occupation and mate choice. Once formed, new polyploids, theoretically, should need to find and occupy a new niche to become established and avoid being out competed by older, presumably more numerous and well-adapted progenitors (Mable, Alexandrou and Taylor, 2011). Historically, polyploids were thought to occur with greater frequency at higher altitudes or latitudes then their diploid progenitors (Löve & Löve 1943). This was thought not only to be a mechanism for finding new niches, but also a way to improve reproductive success by reducing the probabilities of mating between species of different ploidy level, thus increasing reproductive isolation (Otto and Whitton, 2000). However, some evidence suggested that observations of increased environmental range in polyploid species may not be as wide spread as originally thought and/or may be the result of factors other than ploidy level, such as their mode of reproduction (Mable, Alexandrou and Taylor, 2011).

The need for a genetically compatible mate is one of the first obstacles in the establishment of new polyploid lineages (Otto and Whitton, 2000). This issue can be negated by selfing, asexuality and perenniality (long life span). However sexually reproducing organisms face genetic compatibility issues. For example, tetraploids may mate with diploids, which will lead to the formation of triploids. Triploids are frequently considered to be an evolutionary dead-end because they often suffer from low fertility rates and have difficulties with chromosomal pairing during meiosis. As a result they frequently produce aneuploid gametes (Otto and Whitton, 2000). However, triploids may lead to the production of haploid, diploid or triploid gametes although the rate of production would be low. As a result it has been suggested that triploids may facilitate the occurrence of tetraploids at the population level (Ramsey and Schemske, 1998). It has also been suggested that newly formed polyploid species may mate assortatively according to cytotype (cellular characteristics e.g. chromosome number). In animals, such as anurans and fish, mates of similar ploidy level may be selected by using variations in mating call or olfactory cues (Keller and Gerhardt, 2001; Mable, Alexandrou and Taylor, 2011). Furthermore, a number of polyploid plants species have exhibited shifts in flowering time, which would allow them to reduce the chances of fertilisation with plants of different ploidy levels and increasing mating isolation (Thompson and Lumaret, 1992).

An organism's propensity towards polyploidisation is also correlated with a number of key attributes. Mable (2004) observed that the majority of animal polyploids are known to produce large numbers of gametes, both male and female. They also observed that most known, sexually reproducing, animal polyploid lineages also undergo external fertilization.

Therefore they theorised that these characteristics of gamete production and fertilization would facilitate random mixing of gametes. A process which may increase the probability of producing viable polyploid offspring with balanced chromosome sets and also increase the probability of polyspermy (Mable, Alexandrou and Taylor, 2011).

Characteristics associated with the external reproductive environment may also affect the likelihood of polyploidisation through unreduced gamete formation. Polyploid frequencies increase at higher latitudes and altitudes where environmental fluctuations such as temperature are greater this may increase the probabilities of unreduced gamete formations (Mable, 2004). Polyploidy is frequently observed in amphibians and fish, both of whom reproduce in aquatic environments; these environments as breeding grounds for ectotherms, are subject to variability in terms of temperature and potentially pH, which during times of environmental instability may be considerable (Mable, Alexandrou and Taylor, 2011). As a result large numbers of individuals may be exposed to comparatively large temperature (or pH) fluctuations, which may increase the probability of the production of unreduced gametes (Mable, 2004).

1.1.2 Consequences of Polyploidy

The potential evolutionary impact of polyploidy within populations has been vigorously discussed with two opposing arguments coalescing. The first argument hypothesises that polyploidy may represent multiple commonly occurring mutations that may on occasion arise within a population if their phenotypic impact is low. If this were the case then polyploidy would have a negligible role in evolution (Otto and Whitton, 2000). Stebbins argued that while polyploidy had played an important role in the development of some large and widespread genera, its function in the preservation of old genera was greater than its role in the production of new genera (Stebbins, 1940). A view that was later interpreted as regarding polyploids as evolutionary dead-ends (Soltis, Visger and Soltis, 2014). Conversely, the second argument theorises that polyploidy is common within some taxonomic groups because it assists in bringing about faster rates of evolution and provides potentially alternate evolutionary pathways (Otto and Whitton, 2000). Ohno advocated this theory in his book "Evolution by Gene Duplication" (Ohno, 1970 as cited in Furlong & Holland 2004) in which he argued that gene duplication (including polyploidization and tandem gene duplication) was an important mechanism in the evolution of organism complexity (Ohno, 1970 as cited in Furlong & Holland 2004). WGD has also been frequently implicated in the high diversity observed in taxa such as the angiosperms and teleost fish (Masterson, 1994; Peer, Maere and Meyer, 2009; Santini *et al.*, 2009; Pasquier *et al.*, 2016).

Whatever role polyploidy may have in terms of evolution a number of phenotypic and genetic consequences have been observed in polyploid species. One of the most commonly observed phenotypic characteristic in polyploids is that they tend to have a greater cell size then closely related diploids (Cavalier-Smith 1978; Ching et al. 2010; Otto & Whitton 2000). This may then lead to alterations in development/maturation speed and metabolic activity because of alterations in the cell surface area to volume ratio (Otto & Whitton 2000; Weiss et al. 1975). In addition, intracellular distances may change and have implications for signal transduction within the cells (Benfey, 1999). This overall change to cell size would be expected to affect individual size and organ function within polyploid organisms as observed, for example in ovarian retardation in triploid salmon (Benfey, 1999). However, many polyploid species do not show significant differences in whole organism size ranges (Otto and Whitton, 2000). For example, in triploid salmon, although cell size increases, cell number decreases and organ and organism sizes are consistent with sizes recorded for related diploids (Ching *et al.*, 2010).

A frequently recurring though controversial phenotypic observation is that polyploids are often found to have broader ecological tolerances than related diploids (Mable, Alexandrou and Taylor, 2011). This view was based on the (previously discussed) observation that many polyploids seemed to regularly occur in environments that were considered harsher, such as at higher altitudes or more polar latitudes (Löve & Löve 1943). One proposed explanation for this observation is that polyploidy may provide metabolic flexibility, allowing enzymes with shared functions but slightly different forms to be produced, which may each be most effective under different environmental conditions (Otto and Whitton, 2000). However, the observations relating broader ecological ranges to polyploid species are subject to a number of biases and do not appear to be applicable to polyploids in general (Mable, Alexandrou and Taylor, 2011). Thus, the observations of more polyploids in extreme habitats may not be a result of improved ecological tolerances but artefacts of niche partitioning, shifts in mating strategy towards autogamous reproduction and/or the results of a strategy for securing reproductive isolation in successful polyploid species (Otto and Whitton, 2000; Mable, Alexandrou and Taylor, 2011).

Polyploidy also has a number of theoretical and observed consequences at the genetic level. It has been argued that polyploidy can reduce selection efficiency due to the occurrence of multiple alleles at each gene (Otto and Whitton, 2000). So at higher ploidy levels the spread of a beneficial allele may be slower because the selective outcomes of the beneficial allele are diluted by the occurrence of many alternative alleles. Gorelick and Olson argued that polyploidy lineages would show an increase in genetic drift and mutation but with negligible

changes to selection, resulting in non-adaptive radiation (Gorelick and Olson, 2013). A counter argument to this was that organisms with higher ploidy levels would carry more alleles so may be expected to have a higher chance of carrying an allele with a beneficial mutation than organisms of lower ploidy level (Otto and Whitton, 2000). It has also been hypothesised that newly beneficial alleles which were originally deleterious may be present at higher frequencies in organisms with higher ploidy levels. This is because deleterious alleles persist for longer and at higher frequencies in organisms with higher ploidy levels due to the masking effects which also occur with increasing frequency at increased ploidy (Otto and Whitton, 2000).

The influence of ploidy level on beneficial allele spread is intimately associated with the population size and dominance of the allele in question. Otto & Whitton (2000) predicted that organisms with a higher ploidy status would have a fitness advantage over those with lower ploidy status if they were in relatively small to moderately sized populations, and with dominant, or at least partially dominant, beneficial alleles. This prediction was based on the concept that the increases in fitness were less dependent of selective efficiencies and more dependent on how frequently beneficial mutations appear and are established within populations.

Theoretically, the masking of deleterious mutations could provide a temporary advantage to taxa with higher ploidy levels. Otto & Whitton (2000) demonstrated an immediate advantage to organisms that have recently undergone polyploidization because under the right conditions higher polyploids may be better able to mask single copy deleterious mutations then organisms of lower ploidy. However individuals of a higher ploidy level also have a greater chance of bearing a deleterious mutation and these mutations are more likely to persist for longer in organisms with higher ploidy levels (Otto and Goldstein 1992; Otto and Whitton 2000). It is predicted that over time higher polyploids would have a greater deleterious genetic load then taxa with lower ploidy levels but additionally the transitory advantage of higher polyploids may persist for several generations (Otto and Whitton, 2000).

Overall the impacts of polyploidy at the genetic, individual and population level are not clear-cut. There are, however, theoretical treatments demonstrating hypothetical effects of polyploidy at various biological levels, although these are strongly interlinked with other factors such as mode or reproduction, population size and allelic dominance.

1.2 Host parasite interactions and fish immunity

1.2.1 Background of host parasite interactions

A parasite is defined and identified as an organism that lives in or on another organism, known as the host, obtaining nourishment and nutrition from it, causing it a degree of harm and showing a certain level of adaptation to it (Poulin, 2007). Parasites often have highly complex life histories, which can traverse multiple life stages each of which may be associated with a different host (Poulin, 2007). They also range in their host specificities depending on their requirements (Poulin, 2007). A core similarity linking different parasites however is that their transmission cycle terminates in a definitive host which is defined as the host in which sexual maturity is reached and from within which reproduction occurs (Poulin, 2007).

Parasitic species are numerous and are frequently broken down according to their preferred site of residence within or on the host (Jones, 2001). Ectoparasites are those that live outside the host while endoparasites live within the host either as hematozoic parasites (within blood), histozoic parasites (living in host tissue but outside of cells) or coelozoic (living within the host intestinal canal) (Jones, 2001).

A number of factors are associated with a hosts proclivity to harbouring a parasitic infection. These include host age, size, behavioural patterns, physiology, diet, immunology and general condition along with abiotic environmental variables such as temperature (Ryce, Zale and MacConnell, 2004; Khan, 2012; Lester and McVinish, 2016). Hosts also adopt differing strategies for handling parasite infections. These may be broadly broken down into tolerance or resistance strategies where tolerance is defined as the ability to limit the damage caused by a parasite and resistance is the capacity to limit overall parasite burden through immunological mechanisms (Råberg, Graham and Read, 2009).

The overall cumulative effect of the complex life histories of parasite life cycles and the factors that effect a hosts likelihood of carrying an infection, including tolerance vs resistance strategies mean that studying host parasite relationships and communities is a challenging field. There are numerous factors that need to be accounted for in order to get any degree of resolution or clarity over these community wide interactions.

1.2.2 Fish immunity

The immune systems of animal species act as defence mechanisms against a broad array of pathogens. The immune system of fish is not dissimilar to that found in mammals and other higher vertebrates and is made up of both innate and adaptive mechanisms (Alvarez-Pellitero 2008). The innate system is germline encoded and dependent on a wide range of pathogen

recognition receptors (PRRs) with broad specificity (Alvarez-Pellitero, 2008; Medzhitov & Janeway, 2002). Conversely, pathogen recognition in the adaptive immune system is based on antigen receptors with very narrow specificities (Alvarez-Pellitero, 2008). In fish, the adaptive immune system is considered to be intrinsically limited, which means that fish are heavily reliant on the efficiency of their innate immunity (Uribe *et al.*, 2011). These limitations are thought to arise, at least in part, because the poikilothermic nature of fish means that they are subject to environmental temperature fluctuations which has impacts on the rate of physiological functions like enzymatic activity (Uribe *et al.*, 2011). This has been linked to the limited antibody repertoire and slow proliferation, maturation and memory of lymphocytes in fish immune systems (Magnadóttir, 2006; Uribe et al., 2011; Whyte, 2007).

1.2.3 Immune tissues and cellular elements of fish

As with other vertebrates, the immune system of fish is made up of a number of elements including tissues, cells and humoral factors (Alvarez-Pellitero 2008). Fish lack bone marrow and lymph nodes for immune cell development and, as a result, the development of myeloid cells, which include neutrophils and macrophages, occurs primarily in the head kidney (HK) and/or the spleen (Alvarez-Pellitero 2008; Zapata et al. 2006). Primary lymphoid organs in fish include the thymus, kidney and spleen, which are responsible for T cell production, haematopoiesis and antigen phagocytosis respectively (Alvarez-Pellitero, 2008; Zapata et al.,2006).

The first line of immune defence in fish is the physical barrier formed by epidermis, gills and mucosal epithelia (Uribe *et al.*, 2011). Pathogens must overcome these physical barriers to establish host infection. The epidermis is able to respond to pathogenic attack through thickening and cellular hyperplasia (Uribe *et al.*, 2011). In addition, as in general vertebrate immunology, recognition of foreign bodies may also occur through epithelial cells, which may then activate further immune responses (Fritz et al., 2007).

Cellular components of the innate immune systems of fish and other animals include but are not restricted to; phagocytes, pro-inflammatory cells and non-specific cytotoxic cells (Neumann et al. 2001; Alvarez-Pellitero 2008). Phagocytosis is the process by which materials such as cellular debris, microorganisms, macro-molecular aggregates and cells may be ingested into phagosomes (Neumann, et al., 2001). Mammalian eosinophils and mast cells are types of pro-inflammatory cells (Stone et al., 2010). However, in fish the presence or absence of a distinction between eosinophilic granule cells (EGC) and mast cells remains a point of contention (Alvarez-Pellitero, 2008; Reite, 1998; Rocha & Chiarini-Garcia, 2007). Non-specific cytotoxic cells (NCCs) and natural killer (NK)-like cells have also been identified in a number of fish species (Evans et al. 1984a; Evans et al. 1984b). NCCs have a capacity to lyse target cells

via cell-to-cell contact and have similar morphology and function to human NK cells, although NCCs lack cytoplasmic granules present in NK cells (Evans et al., 1984a; Evans et al., 1984b).

The two major cell types forming the adaptive immune system in vertebrates are the T and B cell lymphocytes (Scapigliati, 2013). These two cell types are functionally distinct in that B cells produce and secrete soluble antigen receptors, which are then distributed through the body of an individual (Scapigliati, 2013). Conversely T cells retain antigen receptors on their cell surface and, as a result, interact with foreign materials through direct cell-to-cell contact (Scapigliati, 2013). Distinct B and T lymphocytes have been demonstrated in teleost fish and are thought to be similar in a number of respects to mammalian B and T lymphocytes (DeLuca et al, 1983 as cited in Scapigliati, 2013).

1.2.4 Proteomic mechanisms of fish immunity

In addition to the tissue and cellular components of the fish immune system, a number of protein-based agents have also been documented with regards to the fish immune system (Alvarez-Pellitero 2008). These have roles in the detection of pathogenic material, inter or intra cell signalling, or assist in biostatis or biocide. The factors included in this section comprise pathogen recognition receptors (PRRS), complement, major histocompatibility receptors, antibodies, cytokines, antimicrobial peptides, protease inhibitors, lysozyme and pentraxins.

In fish one of the primary mechanisms of the innate immune system involves the PRRs (Rajendran *et al.*, 2012). PRRs are germ line encoded, have broad specificities and are adapted to recognise conserved pathogen-associated molecular patterns (PAMPS) (Alvarez-Pellitero, 2008). In fish three of the main receptor types include: toll-like receptors (TLRs), nucleotide binding oligomerization domain (NOD)-like receptors (NLRs) and retinoic acid inducible gene I (RIG-I)- like helicases (RLHs) (Aoki and Hirono, 2006; Chang *et al.*, 2011; Rajendran *et al.*, 2012).

TLRs represent a group of type I transmembrane proteins which recognise extracellular PAMPS and initiate innate immune mechanisms when activated (Zhao *et al.*, 2013). TLRs are thought to play a direct role in activation of the innate inflammatory response. They influence adaptive immune responses through regulation of antigen presentation on dentritic cells and through direct effects on T and B lymphocytes and may also induce apoptosis (Salaun et al., 2007). NLRs are intracellular PRRs, which are capable of inducing inflammation and apoptosis in a range of animals (Rajendran *et al.*, 2012). This group of PRRs have been identified via gene and gene expression data in fish species relatively recently (Rajendran *et al.*, 2012). PRRs belonging to the RLH group recognise viral RNA PAMPs in cytoplasmic regions (Chang *et al.*, 2011). These PRRS are thought to have roles in caspase activation, immune signalling and apoptosis (Chang *et al.*, 2011).

1.2.5 Polyploidy and Immunity

Genes that encode immune proteins are often highly polymorphic and it is thought that the high diversity if favoured by a number of pathogen-mediated balancing selection mechanisms (Spurgin and Richardson, 2010; Netea, Wijmenga and O'Neill, 2012; Phillips *et al.*, 2018). There are a number of theories relating to how polyploidy may affect immune efficiency, tolerance and resistance (King, Seppälä and Neiman, 2012). A number of studies have attempted to test these theories and examine the relationship between polyploidy and immunity. King et al. (2012) suggested that polyploidy might increase resistance to parasites for a number of reasons, including: allelic diversity and expression level, which may be increased with ploidy level, along with alterations in physical condition.

The first of the theories put forward by King et al. (2012) noted that parasite-mediated selection should support the persistence of polyploid individuals because of heterozygote advantage and negative frequency dependence. The theory of heterozygote advantage (in terms of immunity related fitness) is based on the principal that individuals heterozygous at pathogen recognition receptor loci (or across replicated loci) will have an advantage over homozygotes because individuals will be able to recognise and respond to a greater range of pathogens or respond more efficiently to a single pathogen due to a greater range of antigens being detected (Spurgin and Richardson, 2010). Heterozygotes have previously been shown to mount more effective immune responses than homozygotes when challenged by pathogens (Doherty & Zinkernagel 1975, Oliver and Piertney, 2012). Polyploid taxa may be expected to have an increased probability of heterozygosity due to the presence of additional haplotypes following genome duplication (King, Seppälä and Neiman, 2012). As a result where heterozygosity in hosts makes evasion of immune recognition systems harder for potential pathogens, polyploidy would be expected to increase resistance in the host (Nuismer and Otto, 2004).

Polyploids may also have an advantage over diploids as a result of negative frequency dependence (rare allele advantage). This hypothesis suggests that pathogens will be under strong selective pressure to infect the most common host genotypes (Koskella and Lively, 2009). This results in common host genotypes having a lower fitness than rare host genotypes. Over longer time scales this host-parasite relationship may become cyclical, with the originally dominant host genotypes becoming rarer and being supplanted by previously rare uninfected host genotypes (Carius, Little and Ebert, 2001; Koskella and Lively, 2009). King et al. (2012) argued that the presence of an additional genome would increase the probability that

polyploids would possess a rare variant within the wider community, which would be advantageous in terms of pathogen resistance.

The dual effects of hybridisation and polyploidisation (allopolyploidy) may also have an effect on host-parasite interactions. The effects of hybridisation could work both in favour of and against the host organism however. The "hybrid-bridge" hypothesis originally proposed by Float & Whitham (1993) argued that plant hybrid intermediates would facilitate host shifting of herbivores between parent species. They also added that this process might also facilitate host shifting in parasites (Float and Whitham, 1993). Conversely however hybridisation may allow inheritance of resistance genes from both parent phenotypes to be expressed conferring greater parasite resistance in hybrids (Jackson and Tinsley, 2003).

Alongside advantages occurring through population dynamics, King et al. (2012) also suggested that differences in protein and mRNA content, relating to immune gene expression, might give polyploids an advantage in resisting parasites. Some data does suggest that transcriptome size increases in polyploids, and positive dosage effects have also been observed where expression increases with ploidy level. However, the relationship between gene expression and ploidy level is complex and patterns may be specific to individual taxa (Otto and Whitton, 2000; Ching *et al.*, 2010; Coate and Doyle, 2010). Moreover, there may be costs associated with higher gene expression. In the yeast species *Saccharomyces cerevisiae* a doubling of gene expression, such as that observed after a duplication event, was heavily selected against (Wagner, 2005). This relationship was thought to be primarily due to the energetic costs associated with increased expression of messenger RNA and protein (Wagner, 2005).

In addition to allelic diversity and expression levels, King et al. (2012) also suggested that polyploidy may positively influence the overall condition of individuals and thus place them in a better position to resist infection. There is little data investigating direct comparisons of physiological condition among polyploids, diploids and haploids, but some studies have compared artificially induced triploid fish and control diploid fish in aquaculture. These studies found physiological difference between triploids and diploids including having reduced reproductive success, disrupted gonadal development and increased cell sizes but reduced cell counts in polyploids (Benfey, 1999). They also observed that artificially induced triploid fish were effected more by stress factors and suboptimal rearing then their diploid counterparts (Vale, 2008). However, triploid Pacific oysters (*Crassostrea gigas*) have been observed to respond less to environmental variability then diploids (Duchemin, Fournier and Auffret, 2007).

In terms of immune specific comparisons further studies have taken place in diploid and triploid *C. gigas,* phagocytosis rates were statistically indistinguishable between different ploidy levels. However, the female triploid phagocytic index was significantly higher than both male triploids and all diploids. In addition immune function in triploid *C. gigas* appeared to be less sensitive to changes in environmental factors (Duchemin, Fournier and Auffret, 2007). In a separate study, triploids of the freshwater snail species *Potamopyrgus antipodarum* had lower concentrations of defensive cells then diploid counterparts (Osnas and Lively, 2006). Triploids *P. antipodarum* were more resistant to parasites from remote locations then diploids, although the authors added that in natural environments the parasites in question would probably overcome the initial triploid resistance (Lively *et al.*, 2004). In Chinook salmon, *Oncorhynchus tshawytscha*, mortality was higher in triploids following a natural outbreak of bacterial kidney disease than in diploids (Ching *et al.*, 2010). However, in a separate experimentally controlled immune challenge no difference was observed in mortality between triploid and diploid salmon, although triploid fish showed reduced performance under stress compared to diploids (Ching *et al.*, 2010). In experimentally challenged diploid and triploid Atlantic salmon, *Salmo salar* initial immune reactions via the alternative complement pathway were similar, although recovery was longer in triploids (Langston, Johnstone and Ellis, 2001). The authors of this study suggested that this increased recovery time may impair the triploid's ability to use complement dependent immune mechanisms. The same study also observed that the hypoferraemic response (a mechanism that denies invading organisms access to host iron reserves which are needed for growth) was also slower to be initiated and recovered in triploids compared to diploids. The observed immune differences between ploidy levels made by Langston et al. (2001) were slight but they may result in reduced efficiency in the triploid immune system. However the authors warned that they could not determine if these differences rendered triploids more susceptible to infection compared to diploids (Langston, Johnstone and Ellis, 2001). The majority of these studies do not favour polyploid individuals however it should be noted that these studies were also conducted on artificially induced triploids, in instances like this immune gene diversity (although not expression levels) would be expected to be the same in both diploids and triploids, on account of the triploids in some instances being the result of a single generation.

A rare example of a study which investigated wild polyploids and pathogens was conducted by Šimková et al. (2013). They compared parasite abundance, species diversity and species richness with major histocompatibility complex (MHC) class II diversity in both gynogenetically reproducing triploids (triploids produced by female only genetic material in which sperm activates the egg but does not fuse with it) and sexually reproducing diploids of

the gibel carp, *Carassius gibelio*. They found that over 50% of triploids expressed one of two common genotypes, whereas all diploids expressed rare genotypes (i.e. genotypes were only shared by one or two individuals). In addition, triploids expressed a greater number of MHC alleles, i.e. two to three MHC class II alleles, whereas diploids tended to express one to two MHC class II alleles. Nucleotide and amino acid diversity were found to be significantly higher in diploids when figures were corrected to consider the number of alleles. When comparing the parasitic species richness in triploids and diploid hosts, parasite richness was significantly higher in triploid fish with the first of the two common genotypes. The abundance of parasites belonging to the taxon *Dactylogyrus* was significantly higher in triploid fish with the second of the two common genotypes. In this particular circumstance it is difficult to separate effects caused by ploidy status and by mode of reproduction. The authors suggest that the majority of these observed effects are related to differences in reproductive strategy not ploidy level, and that the costs of sexual reproduction may be offset by limitations to asexual forms in terms of parasite resistance.

1.3 The Corydoradinae as a study system

The Corydoradinae are a species rich subfamily of Neotropical catfishes (family: Callichthyidae), found throughout the fresh water systems in South America (Bonaparte, 1838; Fuller and Evers, 2005). The Callichthyidae are a family frequently referred to as the armoured catfishes. These are characterised by two longitudinal rows of lateral dermal plates covering the entire length of the body along with two or three barbels extending from the junction of the lips to either side of the mouth (Nijssen, 1970). Within the Callichthydae the Corydoradinae make up approximately 90% of the species, with more than 170 valid species and many additional un-described taxa (Alexandrou and Taylor, 2011).

Phylogenetic analysis has identified 9 sequential lineages within the Corydoradinae based on mitochondrial sequencing data (Alexandrou *et al.*, 2011). A more recent study which used restriction site associate DNA (RAD) markers from across the nuclear genome confirmed the existence of nine lineages but positioned lineage 6 between lineage 8 and 9 as opposed to between lineages 5 and 7 (Marburger *et al.*, 2018). Species within these lineages exhibit highly variable diploid genome sizes, which range from 1 to 8pg, and evidence for a number of WGD events has been identified within their evolutionary history (Oliveira *et al.*, 1992; Alexandrou *et al.*, 2011; Marburger *et al.*, 2018). In addition to this RAD sequencing data also indicated differences in haplotype retention, transposable element (TE) abundance and single nucleotide polymorphism (SNP) read ratios across the nine lineages. Lineage 9 was identified as having significantly greater numbers of haplotypes than any other lineage as well as a

significantly greater TE abundance (along with lineage 7) and SNP read ratio (along with lineage 6) (Marburger *et al.*, 2018). These data were somewhat contradictory and suggested a convoluted evolutionary history across the nine lineages with multiple WGD events. Haplotype diversity data indicated that the oldest WGD event was dated to 54-66MYA and encompassing lineages 2 to 9 while SNP read ratios suggested that this event was dated to 35-44MYA and only included lineages 6 to 9 (Marburger *et al.*, 2018). Discrepancies between data sets were explained as symptoms of post WGD re-diploidisation or genome rearrangements. However both haplotype diversity and SNP read ratios also supported a second WGD event dated 20-30MYA encompassing lineages 6 and 9 (Marburger *et al.*, 2018).

A unique aspect of this subfamily is that mixed communities of up to three species, often from different genetic lineages and with varying genome sizes, have been observed coexisting sympatrically (Alexandrou *et al.*, 2011). Some of the relationships within communities and between species of Corydoradinae have been examined further, including the communal propensity towards Müllerian mimicry and resource partitioning (Alexandrou *et al.*, 2011). A number of mimicry rings have been identified within the Corydoradinae; of those investigated 92% of co-mimics were from evolutionarily distinct lineages (Alexandrou *et al.*, 2011).

The majority of the Corydoradinae are omnivorous detritivores and act as benthic scavengers (Nijssen, 1970). Coexisting species of Corydoradinae frequently exhibit different snout morphologies and analysis of dietary overlap using stable isotopes indicated dietary segregation between species with long and short snouts. This would suggest that coexisting Corydoradinae with different snout morphologies feed at differing trophic levels and are able to avoid resource competition through partitioning (Alexandrou *et al.*, 2011).

Preliminary investigations have explored macro-parasite burdens across fifteen individuals from two Corydoradinae species of varying ploidy level: *Corydoras araguaiaensis* (polyploid) and *Corydoras maculifer* (diploid). On average the diploid species had a macroparasite burden nine times greater than the polyploid (Childerstone & Taylor 2012, unpublished). These differences in parasite burden may have a number of possible explanations. Via mechanisms such as heterozygote advantage, negative frequency dependence or increases in immune gene expression, the polyploid *C. araguaiaensis* be more efficient at resisting parasitic infection. Conversely, the diploid *C. maculifer* may be better able to tolerate parasitic infection, i.e. be able to survive a greater parasite load then *C. araguaiaensis*. If this were the case then sampling strategies in the field may only have collected the polyploid survivors of parasitic infection. Survival of these collected polyploids may be a reflection on a number of things including: rare tolerance within the polyploid

population and age demographic (young polyploids who have had limited exposure to parasites). Dietary preferences may also have had some effect on the differing parasite loads in these Corydoradinae species given that organisms feeding at higher trophic levels are often more vulnerable to parasites (Lafferty, Dobson and Kuris, 2006). What is clear is that the Corydoradinae offers a unique opportunity to examine the effects of polyploidy on parasitic resistance in communities of catfishes, given that coexisting species sharing ecological niches but exhibiting variable ploidy level exist in which all species should have been exposed to the same parasitic community.

1.4 Aims and objectives

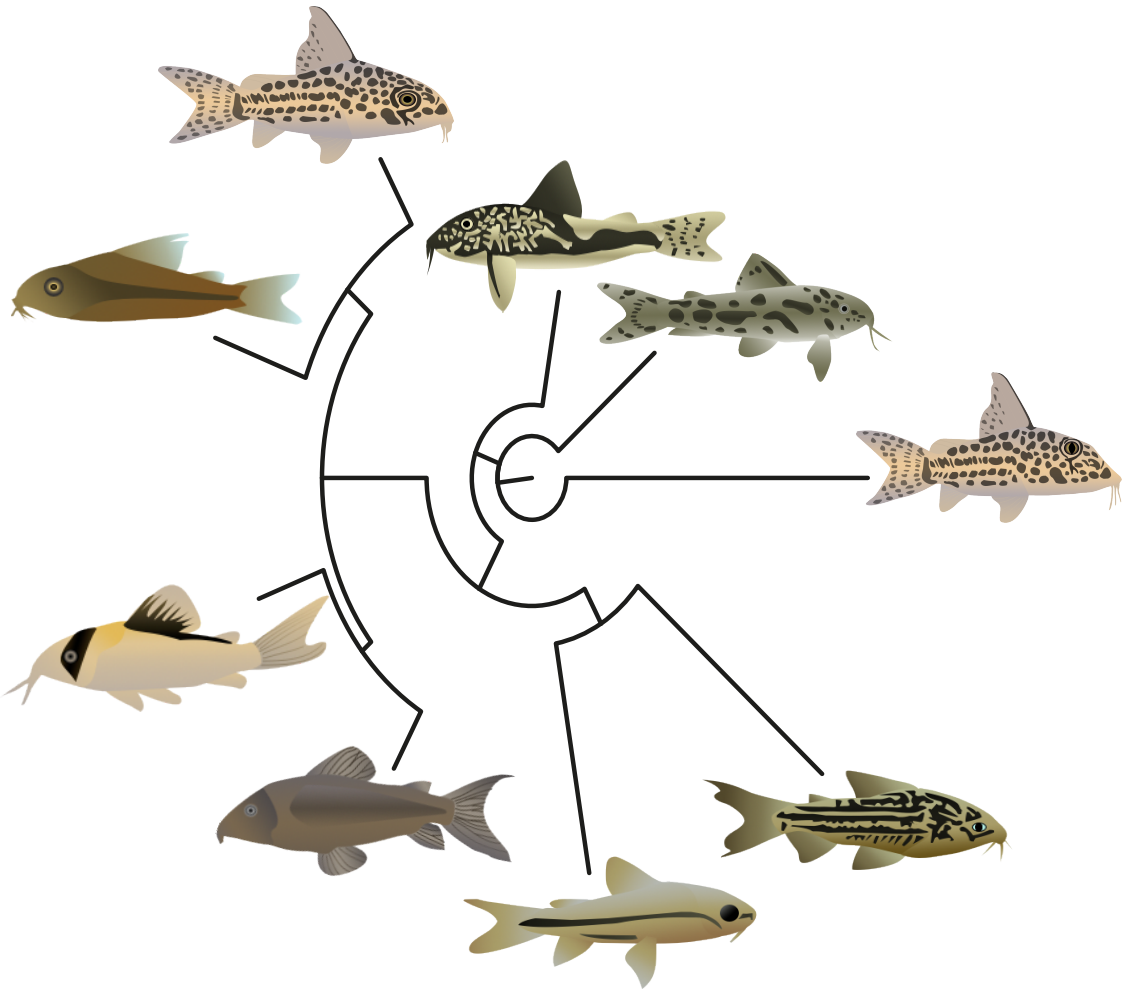
The Corydoradinae present a unique study system for examining direct comparisons of immune gene diversity and parasite load. The Corydoradinae are divided into small sympatric communities of varying genome sizes that should, theoretically, have been exposed to similar pathogenic environments. There are a number of theoretical advantages to maintaining duplicate immune gene copies following a whole genome duplication event, several of which have been described above. The pathogen recognition receptors (PRRS) are a primary mechanism of the innate immune system (Rajendran *et al.*, 2012)and toll-like receptors are one of the three major classes of PRR in fish species (Aoki and Hirono, 2006) making them an excellent starting point for immune gene analysis. The aims of this research are fourfold:

1. To characterise diversity and haplotype retention in a TLR gene across the nine *Corydoradinae* lineages. These lineages have different histories of WGD, it is expected that higher lineages (that have undergone one or two WGD events) would exhibit higher diversity and retain more haplotypes then lower lineages.
2. To characterise a subset of TLRs in a population wide sample of two fish species from the Araguaia river community - *C. maculifer* (a diploid) and *C. araguaiaensis* (a putative tetraploid) - in order to explore relationships between genome size, functional diversity and haplotype retention in immune genes. It is expected that the putative tetraploid, *C. araguaiaensis* would have retained its duplicated TLR copies and that between these four haplotypes there would be a greater degree of diversity then that shown in *C. maculifer*.
3. To examine parasite count and community data from the aforementioned Araguaia River community (consisting of *C. maculifer* and *C. araguaiaensis* species) and compare it to functional immune gene diversity in the already characterised TLRs. Preliminary data suggested that *C. maculifer* harboured a greater parasite burden the *C. araguaiaensis*. We would expect this earlier trend to hold true although whether this

might be linked to tolerance in *C. maculifer* or resistance in *C. araguaiaensis* remains to be seen.

4. To characterise and explore variation in the PRR gene family across the *C. maculifer* genome. Information regarding immune genes is negligible in the Corydoradinae. Identifying and characterising the PRR gene family across this species genome will facilitate further understanding of immune genetics in these fish.

Chapter 2:
Characterisation of Toll-like Receptor 2 across the nine *Corydoras* Lineages



2.1 Introduction

Whole genome duplication (WGD) events, i.e. events which result in the duplication of all genetic material at least once, have been observed in a number of taxa although these are largely plants and ectothermic animals (Mable, Alexandrou and Taylor, 2011). WGD events are thought to be linked to increased genetic diversity and may be a mechanism for enabling adaptive radiations and genetic innovation (Mable, Alexandrou and Taylor, 2011). However genome rearrangement and gene fractionation (a mechanism of re-diploidisation) frequently follow a WGD event resulting in much of the duplicated genetic material being rapidly lost (Berthelot *et al.*, 2014). In some gene families, such as the immune genes, there are theoretical advantages in retaining additional gene copies (haplotypes) despite the on-going re-diploidisation processes (King, Seppälä and Neiman, 2012).

The subfamily Corydoradinae are a highly species rich group of Neotropical armoured catfishes. Species belonging to this subfamily are distributed throughout fresh water catchments of South America, live in mixed sympatric communities and have a convoluted evolutionary history of genome expansion via whole genome duplication events (Bonaparte, 1838; Fuller and Evers, 2005; Alexandrou *et al.*, 2011). A relatively recent set of phylogenetic analyses based on mitochondrial data split the Corydoradinae into nine major lineages (Alexandrou *et al.*, 2011). This, in addition to flow cytometry analysis, which indicated highly variable diploid genome sizes ranging from 1pg to 8pgs across the nine lineages, suggested multiple WGD events in the evolutionary history of the Corydoradinae (Oliveira *et al.*, 1992; Alexandrou *et al.*, 2011). These inferences were further developed using Restriction site Associated DNA (RAD) sequence data, which provided representative proportions of comparable genomic data for a subset of individuals for each lineage. Phylogenetic inferences from these data largely supported those derived from the earlier mitochondrial data with one exception. RAD sequence data showed lineage six clustering with lineage nine instead of between lineage 5 and lineage 7 (Marburger *et al.*, 2018). In addition RAD data indicated that species from lineage 9 had markedly higher haplotype retentions (preservation of additional gene copies) per contig, transposable element abundances and SNP read ratios per contig (Marburger *et al.*, 2018).

The Toll-like Receptor (TLR) gene family are highly polymorphic, have an evolutionary history associated with duplication events and lineage specific gene loss or gain (Hughes and Piontkivska, 2008; Netea, Wijmenga and O'Neill, 2012; Solbakken *et al.*, 2018). Phylogenetic evidence suggests that vertebrate TLRs incorporate two ancient groups that are thought to have arisen through gene duplication events before protostomes and deuterostomes diverged (Hughes and Piontkivska, 2008). This phylogenetic system groups mammalian TLR1, TLR2, TLR6

and TLR10 into a single unit and the remaining mammalian TLRs into a second unit, and observes that functional similarity is maintained within these units (Hughes and Piontkivska, 2008). Functional specialisations are thought to have arisen and been maintained following gene duplications within the ancestors of these units (Hughes and Piontkivska, 2008). Phylogenetic analysis suggests that six TLR genes are shared between fish, mammals and birds (Temperley *et al.*, 2008). This analysis also indicated that TLR1-like genes arose independently in fish, mammals and birds from a common ancestor, while the remaining TLRs were already present prior to the splitting of the major vertebrate lineages, and any of those now missing within specific lineages have been lost subsequently (Temperley *et al.*, 2008). In Gadiformes, these ancient genome expansions correlate with a loss in the major histocompatibility complex class II (MHCIII), and evidence from selection analyses suggests that this loss might have encouraged the development of new TLR innovations within this Order (Solbakken *et al.*, 2018).

2.1.1 Aims and objectives

Here, we characterise TLR2 structure across the nine *Corydoras* lineages, compare diversity and haplotype retention in this immune gene and reference the outcomes from this analysis back to the broader diversity and haplotype retention found in earlier RAD sequence data (Marburger *et al.*, 2018). Both the species sub-set and the immune gene examined here have convoluted evolutionary histories regarding duplication events, however we would predict that haplotypes of immune genes are more likely to be retained over other genes because of the potential advantages this might incur. In addition to previously acquired RAD data from across the nine Corydoradinae lineages, sequence data for TLR2 were collected from a single individual representative of each of the nine lineages. An overview of the species included in this chapter, their geographic distribution and the data available for each species is represented in Table 2.1.

Table 2.1: A summary of lineage, sequence data available and the natural geographical distribution of the Corydoradinae species analysed within this chapter.

Species	Lineage[1]	Sequencing data available	Distribution[2]
Corydoras maculifer	1	TLR1/TLR2	Brazil; Est. Mato Grosso, Rio das Mortes, Rio Araguaia
Corydoras fowleri	1	RAD	Brazil, Peru, Colombia
Aspidoras poecilus	2	TLR2/RAD	Brazil; Est. Mato Grosso, upstream of Poroi village, Rio Xingu, Rio Araguaia
Scleromystax kronei	3	TLR2/RAD	Brazil; Sao Paulo, Rio Betari, Rio Iporanga, Rio Juquia
Corydoras pygmaeus	4	TLR2/RAD	Brazil; Est. Rondonia, Rio Madeira, Ecuador; Napo Province, Rio Aguarico, Peru; Loreto Province, Rio Nanay
Corydoras elegans	5	TLR2/RAD	Brazil; Est. Amazonas, Rio Amazonas, Peruvian and Colombian Amazon, Ecuador; Rio Aguarico, Rio Napo
Corydoras nattereri	6	TLR2/RAD	Brazil; Rio de Janeiro, Rio Paraiba do Sul drainage, Sao Paulo, Rio Juquia
Corydoras aeneus	7	TLR2/RAD	Western Trinidad, Argentina; Rio Parana, Paraguay; Rio Paraguai, Bolivia; Rio Itenez, Colombia, Brazil; Est. Rondonia, Venezuela; Rio Apure, Guyana region
Corydoras imitator	8	TLR2/RAD	Brazil; Est. Amazonas, Rio Negro
Corydoras araguaiaensis	9	TLR1/TLR2/RAD	Brazil, Est. Mato Grosso, Rio Araguaia

1- Marburger et al. 2018, 2 – Fuller & Evers 2005

2.2 Methods

2.2.1 Sampling and DNA extraction

Individuals of *Corydoras maculifer* (n=17) and *Corydoras araguaiaensis* (n=36) were collected from the wild from the same location in the Araguaia region in Brazil by Martin Taylor (MIT), Claudio Oliveira (CO) (2012 and 2015) and Ellen Bell (EB) (2015), euthanised by anaesthetic overdose and stored individually in 100% ethanol. Single individuals from the remaining seven *Corydoras* lineages (*Aspidoras poecilus, Scleromystax kronei, Corydoras pygmaeus, Corydoras elegans, Corydoras nattereri, Corydoras aeneus* and *Corydoras imitator*) were collected between 2005 and 2013 by MIT and CO and stored as above. DNA was extracted from fin clip tissue using the salt extraction protocol after Sunnucks & Hales (1996) and Aljanabi & Martinez (1997).

2.2.2 PCR amplification, library preparation and sequencing

Prior to PCR primer design the complete gene (complete coding sequence (CDS)) for TLR1 and TLR2 in *Ictalurus punctatus* (channel catfish) were downloaded from the National Centre for Biotechnology Information (NCBI) genbank and blasted (BLASTn, NCBI-2.2.29) against assembled transcriptome data (MIT unpublished) for an assortment of *Corydoras* species (including: *Corydoras haraldschultzi, Corydoras paleatus, C. aeneus, Corydoras melini, Corydoras cruziensis, Corydoras schwartzi, C. elegans, C. nattereri, Aspidoras fuscoguttatus, Corydoras julii, Scleromystax prionotos, Corydoras fowleri, C. nattereri* and *Corydoras mamore*). Blast outputs were filtered to only include contigs of greater then 100bp in length and more than 70% similarity to the *I. punctatus* TLRs. Matching contigs were extracted from the datasets and aligned to the respective *I. punctatus* TLR sequence (MUSCLE aligner within Geneious-9.0.5). Specific PCR primers were designed, based on conserved regions within the alignments, to amplify *c.* 2.5kb fragments of both TLR1 and TLR2. Primer forward/reverse pair compatibility (i.e. self-complement, Tm and % GC) was checked using Primer3 (within Geneious-9.0.5). Although these first sets of primers (TLR1_univ Fw/Rv and TLR2_univ Fw/Rv) worked well for TLR1 and TLR2 in *C. araguaiaensis,* and for TLR2 in 7 other species of *Corydoras*, they failed to work with *C. maculifer*. As a result TLR1 and TLR2 sequences from *I. punctatus* were blasted (BLASTn, NCBI-2.2.29) directly against the *C. maculifer* genome (MIT unpublished) and specific primers (Mac_TLR1 Fw/Rv and Mac_TLR2 Fw/Rv) designed as before (see Table 2.2). The primers TLR2_univ Fw/Rv were found to work well for the all-remaining *Corydoras* species tested but TLR1 primers could not be optimised sufficiently for use on *Corydoras* lineages 2 to 8. As a result, although the principal focus of this chapter is TLR2, the

processes behind sequencing of both TLRs are closely intertwined, so for the purposes of methodological clarity both TLR1 and TLR2 will be included in this section, but for the resulting analysis of TLR1 see Chapter 3.

For PCR amplification 1.25µl of 10µmol forward and reverse primer, 12.5µl of PCR Master Mix (Phusion High Fidelity PCR Master Mix with HF Buffer) and 2µl of extracted DNA were combined and made up to a final volume of 25µl with H_2O. PCR conditions were: initial denaturation of 98°C for 30s and then a secondary denaturation of 98°C for 10s, species/primer specific annealing temperatures (see Table 2.3) for 30s, extension at 72°C for 120s for 35 cycles with a final extension step at 72°C for 5 minutes. PCR products were visualised on ethidium bromide stained 1.2% agarose gels.

Table 2.2: Primers used to amplify TLR2 across Corydoras samples

Primer name	Forward	Reverse
TLR1_univ	TGGCGATCCTGGTGGCCA	CTCTGCTTGGAGTGCTGCT
TLR2_univ	GCCAGCAGGATCTAAGCGAC	TCGTCCCTTTTTAGAGCGGCC
Mac_TLR1 (*C. maculifer* specific)	AGGATTCACTGGCTATTCTGGAGG	GCAATGGGGTTTGGTAAATCTCG
Mac_TLR2 (*C. maculifer* specific)	GACATTGAGATCATTAGCCAGCAG	CGGCTCTCAGATTGTTCCAGAA

Table 2.3: Species and primer specific PCR annealing temperatures

Species	Primer	Annealing temperature (°C)
C. maculifer	Mac_TLR1	63.0
C. maculifer	Mac_TLR2	68.2
A. poecilus	TLR2_univ	67.0
S. kronei	TLR2_univ	65.2
C. pygmaeus	TLR2_univ	67.0
C. elegans	TLR2_univ	67.0
C. nattereri	TLR2_univ	67.0
C. aeneus	TLR2_univ	67.0
C. imitator	TLR2_univ	65.2
C. araguaiaensis	TLR1_univ	69.8
C. araguaiaensis	TLR2_univ	69.8

2.2.3 Library preparation and sequencing

Two loci were sequenced together for both *C. maculifer* and *C. araguaiaensis* TLR1 and TLR2, however only the TLR2 loci was sequenced for the remaining *Corydoras* species. Two 25µl PCRs were conducted for each locus to ensure sufficient PCR product for library preparation and sequencing. Amplification products for each locus were then pooled (2 X 300ng). Giving two replicates with 600ngs of pooled loci PCR product per individual. Amplicon DNA was fragmented using the NEB ds Fragmentase kit, combining 16µl of pooled DNA, 2µl of 10X Fragmentase reaction buffer V2 and 2µl of ds DNA Fragmentase mixture (diluted 1:10 with buffer). Samples were incubated at 37°C for 17 minutes. Finally 5µl of 0.5M EDTA was added to each reaction to stop enzyme activity and replicate samples were pooled. Fragmented amplicons were cleaned using an AMPure XP bead clean-up kit (Agencourt). A 1.5X bead to product volume ratio was used, and samples were re-suspended in 17µl of H_2O. DNA end repair was then performed using NEBNext End Repair Module, followed by another bead clean-up step at 1.5X bead to product volume, and final re-suspension in 28µl of H_2O. Samples were quantified again using a Qubit fluorometer (Qubit dsDNA HS Assay Kit).

Barcode mixes were made up from sets of partially complementary oligonucleotide sequences with additional modifications, including phosphorothioation (noted by *) to increase stability at the 5' and 3' ends and addition of a 5' phosphate in the multiplex barcode sequences (table 2). Barcode mixtures were diluted to 40µM and contained: 20µl of each complementary 200µM oligonucleotide solution (e.g. GCATG 1 and GCATG multiplex 1), 10µl of 10X annealing buffer (100mM Tris HCL pH 8, 500mM NaCl, 10mM EDTA) and 50µl of nuclease free water. This solution was incubated at 97°C for 2.5 minutes then cooled at a rate of 3°C per minute to 21°C. Barcode mixtures were diluted 1:10 for subsequent use.

Samples (n=60; 53 *C. maculifer* and *C. araguaiaensis* and single individuals from seven other species) were pooled into 12 groups of five individuals for the first round of barcode ligation (Figure 2.1; Step A). Within each group of five, individual DNA extracts were adjusted to match the lowest concentration sample in 23µl. Samples then underwent an A-tailing step using: 1.8µl of NEB Klenow fragment (3' to 5' exo-), 3µl of NEB2 buffer and 10µl of 10mM dATP made up to a total volume of 30µl with H_2O and incubated at 37°C for 30 minutes. This step was immediately followed by barcode ligation; 2.25µl T4 DNA ligase was mixed with 4µl 10mM ATP, 2µl of 50mM MgCl2, 0.2µl of 4µM barcode mix (table 2.4) and made up to 10µl total volume with H_2O before being added to each respective A-tailed sample. Samples were incubated at 16°C for 30 minutes and then at 65°C for 10 minutes before being cooled to room temperature with 2°C decreases every 2 minutes.

Inline barcodes (pools of five samples with differing barcodes) were then combined to produce 12 library pools. Ampure XP beads were used in a size selection step; by altering the ratio of beads to product it is possible to select for larger or smaller fragment sizes. This process was optimised to select for fragment sizes of 700-800bps in length over three bead clean-up steps. The first step was at a bead-to-product ratio of 1:1 and product re-suspension in 50µl, the second and third steps both used 0.8:1 bead to product ratios and re-suspension in 40µl and 15µl respectively. Products were then quantified using Qubit fluorometer (Qubit dsDNA HS Assay Kit).

Table 2.4: Library adaptors and barcodes used to for NextSeq sequencing and sample identification

Ligation Barcode	**Sequence**[1]
GCATG 1	A*CACTCTTTCCCTACACGACGCTCTTCCGATCTGCATG*T
AACCA 2	A*CACTCTTTCCCTACACGACGCTCTTCCGATCTAACCA*T
CGATC 3	A*CACTCTTTCCCTACACGACGCTCTTCCGATCTCGATC*T
TCGAT 4	A*CACTCTTTCCCTACACGACGCTCTTCCGATCTTCGAT*T
CTTGG 18	A*CACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGG*T
GCATG multiplex 1	/5Phos/CATGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCA*C
AACCA multiplex 2	/5Phos/TGGTTAGATCGGAAGAGCACACGTCTGAACTCCAGTCA*C
CGATC multiplex 3	/5Phos/GATCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCA*C
TCGAT multiplex 4	/5Phos/ATCGAAGATCGGAAGAGCACACGTCTGAACTCCAGTCA*C
CTTGG multiplex 18	/5Phos/CCAAGAGATCGGAAGAGCACACGTCTGAACTCCAGTCA*C

PCR Index	**Sequence**
Common PCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
Multi PCR2 Index 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTG
Multi PCR2 Index 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTG

1 Primer sequences adapted from Peterson et al. 2012

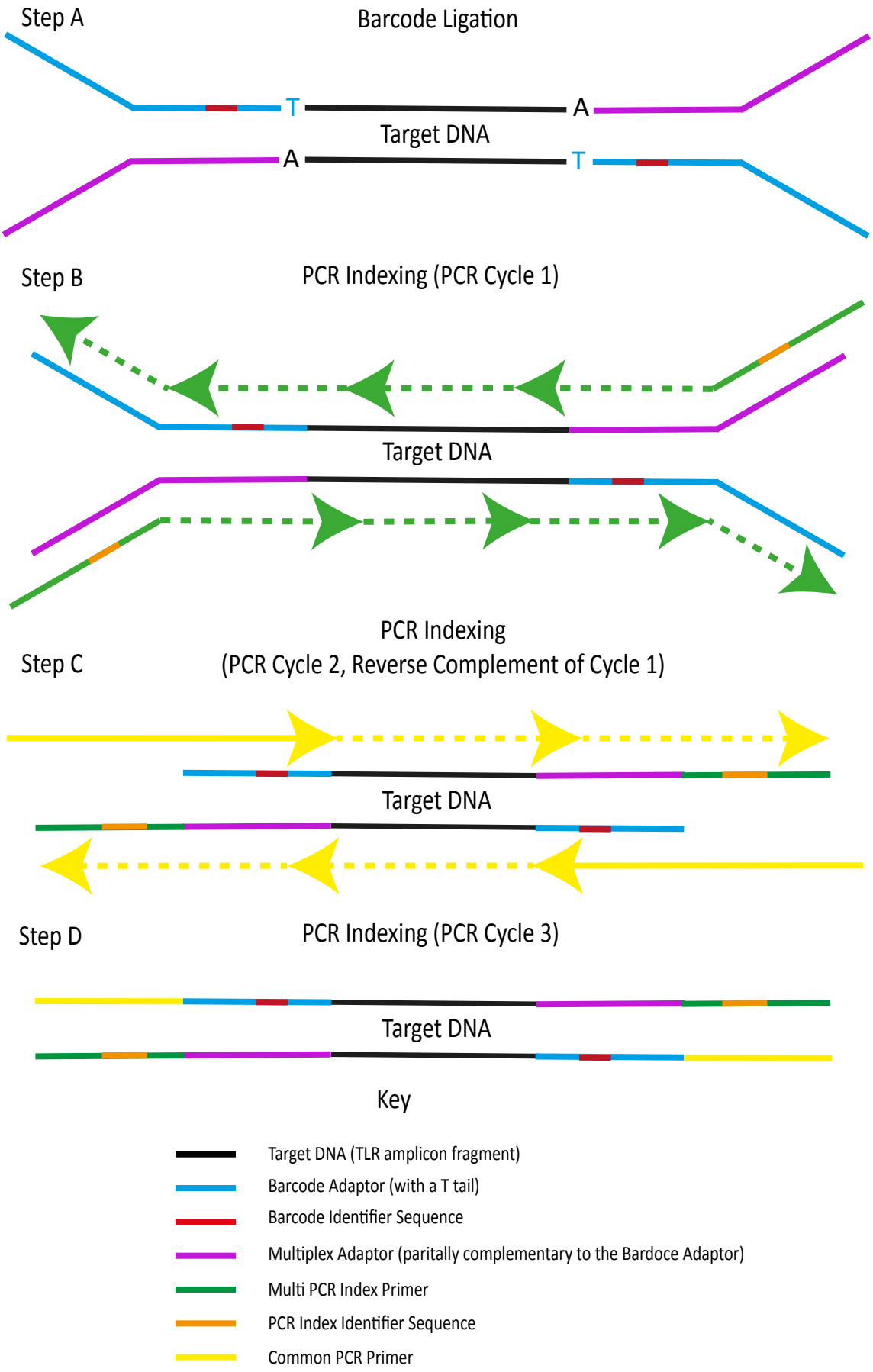


Figure 2.1: The library preparation process including dual barcoding with barcode ligation and PCR annealed indices (not to scale).

Libraries were PCR amplified and indexed (Figure 2.1; Step B and Step C) using 12 different index sequences built into the PCR primers. PCRs were composed of 1µl of pooled library, 10µl of PCR Master Mix (Phusion High Fidelity PCR Master Mix with HF Buffer), 0.2µl of 10µM common PCR1 primer and 0.2µl of 10µM multi PCR2 primer made up to a final volume of 20µl with H_2O. Primers were designed so that on the first PCR cycle only the Multi PCR2 primer would anneal (Figure 1; Step B) and on all subsequent cycles both common PCR1 and multi PCR2 primers annealed (table 2, Figure 1; Step C). The Multi PCR2 primers carried the indices and this ensured that indices were effectively incorporated into the PCR product (Figure 1; Step 4). PCR cycling conditions included: an initial denaturation step at 98^{o}C for 30 seconds, followed by 10 cycles of a denaturation step of 98^{o}C for 40 seconds, an annealing step of 65^{o}C for 30 seconds and an extension step of 72^{o}C for 30 seconds followed by a final extension step at 72^{o}C for 5 minutes. Five PCRs were performed for each of the 12 library pools and inline products were pooled together afterwards. Libraries were then cleaned using Ampure XP beads at a 0.8:1 bead to product ratio and re-suspended at 15µl. Library pools were then quantified using a Qubit dsDNA HS Assay kit and size distributions verified using a Bioanalyzer (Agilent 2100 Bioanalyzer). The 12 library pools were then pooled at equal concentrations into a single library pool quantified once more using Qubit fluorometer (Qubit dsDNA HS Assay Kit) and sent for sequencing on a NextSeq platform.

2.2.4 Data processing and analysis

All sequencing data was quality checked using FastQC (version 0.11.5) and de-multiplexed by ligation index and then by PCR barcode using in-house BASH scripts. Sequence data were then run through Cutadapt (version 1.13) and Trimmomatic (version 0.2.36) to remove traces of adaptor contamination.

Reads from a single individual, selected for its high number of reads, from each species were mapped to TLR1 and TLR2 sequences from the *C. maculifer* genome (MIT unpublished) data using BWA-mem (version 0.7.12, Li & Durbin 2009) and a consensus sequences produced using Geneious-9.0.5. These consensus sequences were edited manually to replace ambiguous bases with non-ambiguous base codes and were used as species specific references for subsequent mapping of raw reads and SNP calling. Reads from all individuals of matching species were mapped to these single individual derived consensus sequences using BWA-mem. SNP calling across populations of *C. maculifer* and *C. araguaiaensis* was completed with FreeBayes (version 1.1.0, (Garrison & Marth 2012)) and filtered to only include SNPs with a minimum of 5 read counts per allele, and that occupied a minimum of 10% of the overall read depth at each site. FreeBayes is optimised for population wide SNP calling and is less precise

with single individuals, therefore for the remaining seven lineages represented by only a single individual, a different methodology was adopted. QualitySNPng (Nijveen *et al.*, 2013) was used for both SNP calling and haplotype counting using short range phasing, and the results of both analyses were validated manually. When SNP calling QualitySNPng was configured to require a minimum number of 5 read counts per allele and a minimum of 10% of overall read depth per allele. All data points were plotted using ggplot2 (version 2.2.1) within R studio (R version 3.4.1).

Haplotype counts were estimated firstly using manually validated QualitySNPng estimates and secondly using the proportional frequencies of SNP read depth. All of these metrics were calculated across a single TLR gene (TLR2) in single representatives of lineages 2 to 8. In the case of lineages 1 and 9 (*C. maculifer* and *C. araguaiaensis*) both haplotype estimates were based on both TLR1 and TLR2 population wide frequencies. Proportional SNP frequencies were plotted in histograms using ggplot2 (version 2.2.1) within R studio (R version 3.4.1) and haplotype estimates from QualitySNPng were fed into downstream phylogenetic analysis.

To ensure that the correct genes had been sequenced, phylogenetic trees were built from protein alignments of both TLR1 and TLR2 loci within the Corydoradinae subfamily and downloaded amino acid sequences from all known TLRs of *Danio rerio* (zebra fish) and *Ictalurus punctatus* (Channel catfish) (Genbank). Trees were also used to look at phylogenetic positioning of *Corydoras* lineages based on nucleotide alignments of TLR2. Maximum likelihood trees, built using IQ-TREE (version 1.5.5, Nguyen et al. 2015; Hoang et al. 2018), were based on the best model fit identified by jModelTest and using the Bayesian information criterion. Trees were visualised using FigTree (version 1.4.3) and a colour overlay based on QualitySNPng minimum haplotype estimates were added to one of the trees using Phytools (version 0.6-44, Revell 2012).

The Simple Modular Architecture Research Tool (SMART, Letunic & Bork 2018) was used to identify and estimate positions of different protein domains based on the translated sequence data of each species TLR2 consensus sequence.

Counts and placement of SNPs shared between the nine lineages were quantified across TLR2 sequence data, and more broadly across already published RAD sequence data (Marburger *et al.*, 2018). For TLR2, SNPs were considered shared if the substitution and location of the SNP was identical between one or more species. These shared sites were enumerated and plotted in a heat map using the ggplot package in R Studio. Shared SNPs were also plotted by position to ascertain if the TLR region was important. To investigate whether patterns at TLR2 were representative of the genome as a whole RAD-seq data generated by

Marburger (2015) were also analysed. RAD sequencing data were cleaned, demultiplexed, assembled (using Velvet), mapped (using BWA-mem) and variants called (using FreeBayes) across two individuals from each of the nine lineages (including: *C. fowleri, A. poecilus, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator* and *C. araguaiaensis*) by Marburger et al. 2018. SNPs from this analysis were further filtered to ensure that only heterozygous, bi-allelic SNPs, at sites with an overall minimum read depth of 10 and minimum presence of 10% of the overall read depth were included in downstream analysis. SNPs shared across two of more lineages were then enumerated and plotted as a heat map using ggplot in R Studio.

2.3 Results

2.3.1 Sequencing, Data Cleaning and Quality Control

The sequencing run produced 876,570 single reads (GC content = 46%) which once filtered, cleaned and trimmed left 762,107 single reads (GC content = 47%). Once de-multiplexed, read counts ranged from 972 to 65726 per library, with read depths per individual ranging from 19.2 to 152.2. A single *C. araguaiaensis* individual (Idx08_Bc2_S2) was removed from downstream analysis due to poor sequencing depth and coverage (total read count of 292 and average depth of 8.4). All other libraries were deemed adequate for further analysis (mapped read count >500 and mean depth >15). Reads retained at each of the clean-up and mapping stages are listed in Table 2.5 along with mean read depth following mapping. Statistics reported for *C. maculifer* and *C. araguaiaensis* are averaged across multiple individuals, while a single individual represents the remaining seven lineages. Read retention rates were fairly uniform between the different bioinformatics stages, with the exception of the *C. imitator* library which had relatively poor read retention between the de-multiplexing and mapping steps. This was thought to be due to multiple band amplification at the initial PCR stage, *C. imitator* proved difficult to optimise and as a result reactions for this individual were untidy. However, sequence data for additional bands would have been removed at the mapping stage of the analysis so shouldn't affect downstream data processing.

To ensure that the identity of targeted TLRs was correct consensus TLR sequences were translated and aligned to complete TLR data sets from *Danio rerio* and *Ictalurus punctatus* prior to phylogenetic analysis (Figure 2.2). Tree topology grouped TLR1 and TLRs from *Corydoras* species with the corresponding TLRs in *D. rerio* and *I. punctatus* suggesting that the correct target loci were sequenced.

2.3.2 Variant calling

Variant data for *C. maculfier* and *C. araguaiaensis* were averaged across their separate populations. SNP counts varied between lineages and are displayed according to substitution type in Figure 2.3. *C. maculifer* (lineage 1) showed the least number of SNPs across the TLR2 gene. SNP abundance was highest in *C. aeneus* (lineage 7) and *C. poecilus* (lineage 2). *C. aeneus* was also the only lineage to show evidence of tri-allelic non-synonymous SNPs (Figure 2.3).

Table 2.5: Sequence read retrieval from the single ended sequencing run

Species	Lineage	Number of loci sequenced	De-multiplexed reads	Mapped reads	Mean depth
C. maculifer (n=17)	1	2	6689 (SD 11096)	2004 (SD 910)	62.3 (SD 28)
C. araguaiaensis (n=36)	9	2	3951 (SD 1223)	3337 (SD 732)	103.5 (SD 23)
A. poecilus (n=1)	2	1	3209	2046	133.4(SD 27.5)
S. kronei (n=1)	3	1	2042	1732	113.8 (SD 25.2)
C. pygmaeus (n=1)	4	1	1682	1264	83.0 (SD 22.5)
C. elegans (n=1)	5	1	2041	1256	82.2 (SD 19.7)
C. nattereri (n=1)	6	1	3804	1333	81.6 (SD 25.4)
C. aeneus (n=1)	7	1	1543	1285	83.9 (SD 18.2)
C. imitator (n=1)	8	1	11490	631	41.1 (SD 12.1)

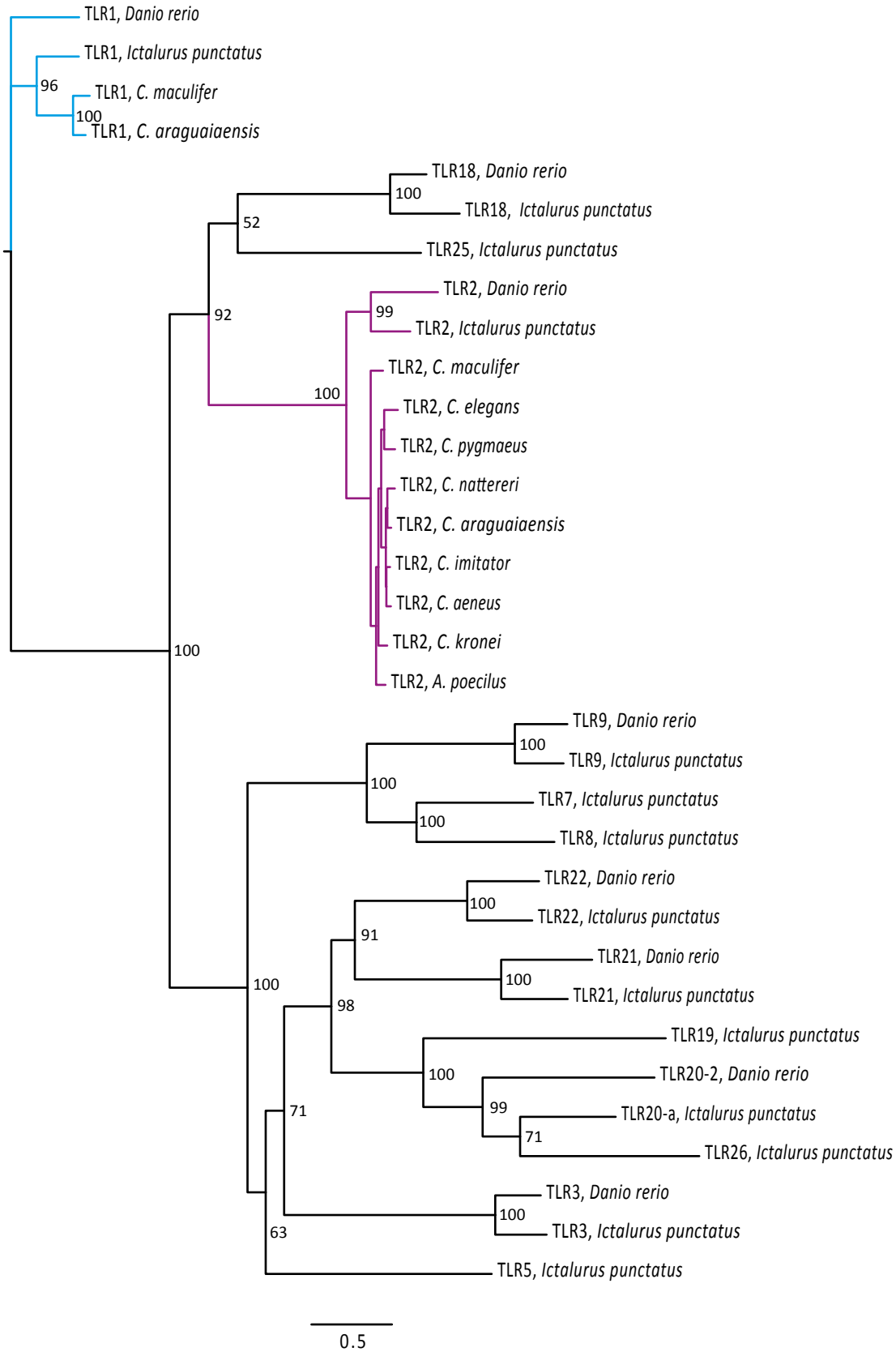


Figure 2.2: Topology recovered from the phylogenetic analysis of protein alignments from TLR1 in C. maculifer and C. araguaiaensis, TLR2 across nine Corydoras lineages, and all known TLRs in D. rerio and I. punctatus. The tree was built in IQ-TREE utilising WAG+F+I+G4 model and left unrooted. Figures at nodes represent bootstrap support. The tree is coloured according to Corydoras TLR1 and TLR2 locations.

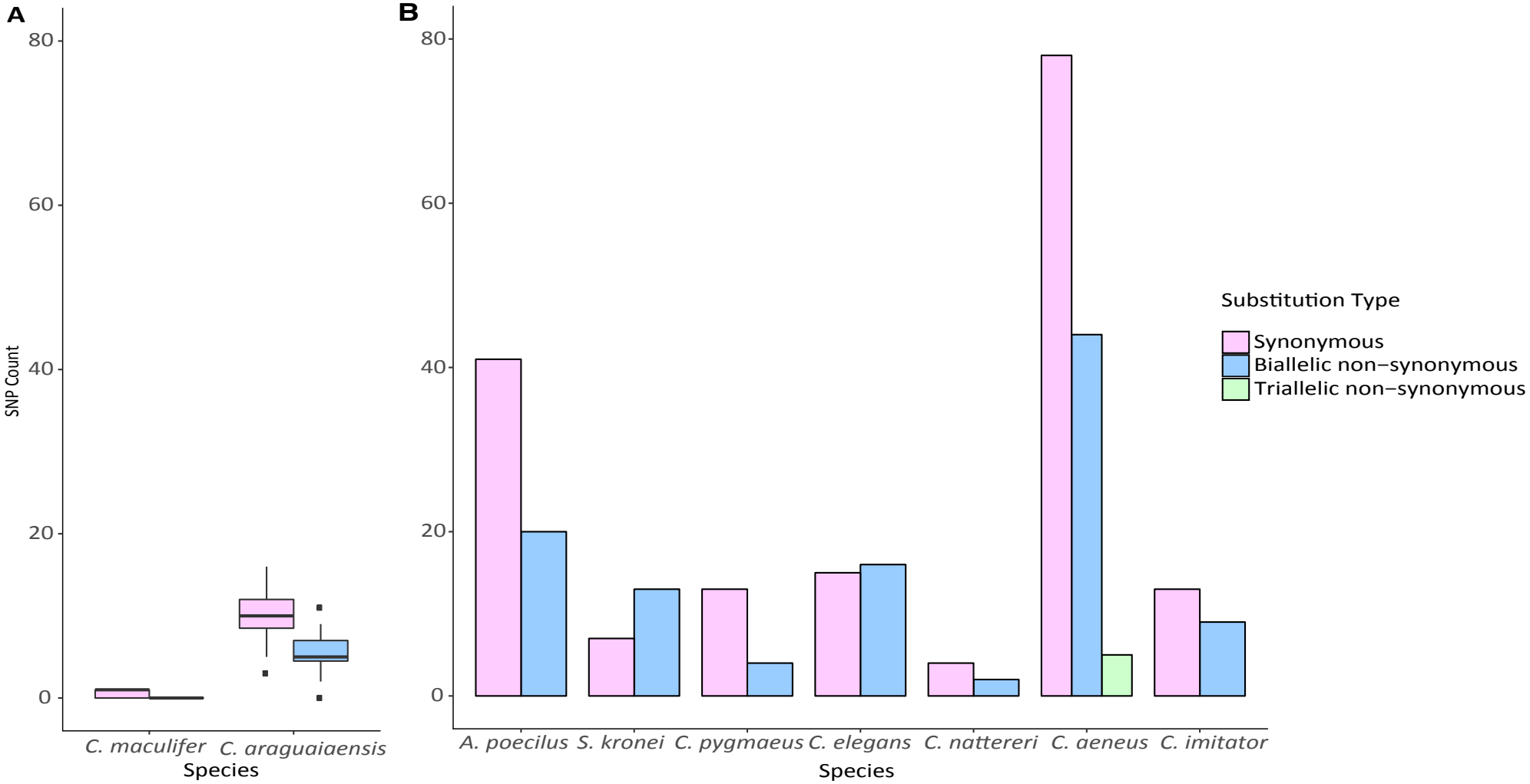


Figure 2.3: TLR2 SNP counts across the nine Corydoradinae lineages. Plot A shows average SNP counts across populations of C. maculfer (n= 17) and C. araguaiaensis (n=35) and Plot B shows total counts across single individuals of each species displayed.

2.3.3 Haplotype number

Phylogenetic analysis (Figure 2.4) based on maximum likelihood for TLR2 showed topologies matching those based on RAD sequencing data (Marburger *et al.*, 2018) but differing from earlier trees built on mtDNA data (Alexandrou *et al.*, 2011). Trees based on mtDNA sequence data placed lineage 6 between lineages 5 and 7 whereas trees based on TLR2 and RAD data place lineage 6 between lineages 8 and 9.

Two methods were used to estimate haplotype retention (or copy preservation) levels of TLR2 for each representative of each lineage. However, the results produced from these two methods did not generally overlap. QualitySNPng suggested that only two species had at least two copies of TLR2 (the diploid *C. maculifer* (lineage 1) and putative polyploid *C. nattereri* (lineage 6)), the remaining lineages had at least three haplotypes, and in the case of *C. aeneus* a signature for ten alternative haplotypes was identified (Figure 2.4). In contrast the SNP read ratio histograms (Figure 2.5), suggested that some SNPs were present at frequencies of 0.10-0.15, which would indicate the presence of 6-10 haplotypes in Lineages 2, 3, 4, 5, 7 and 8. Admittedly strong peaks at frequencies of 0.5 were also identified in lineages 3, 6 and 7 indicating that most SNPs are present in half of the sequenced haplotypes. In the case of *C. nattereri* (lineage 6) the lack of any other peaks supported the outcome from the short range SNP phasing analysis and indicated the presence of two haplotypes. It is worth remembering at this stage that this data is from a single gene in a single individual for seven of the lineages reported, thus results are only suggestive, and no firm conclusions on more general population wide haplotype counts can be made.

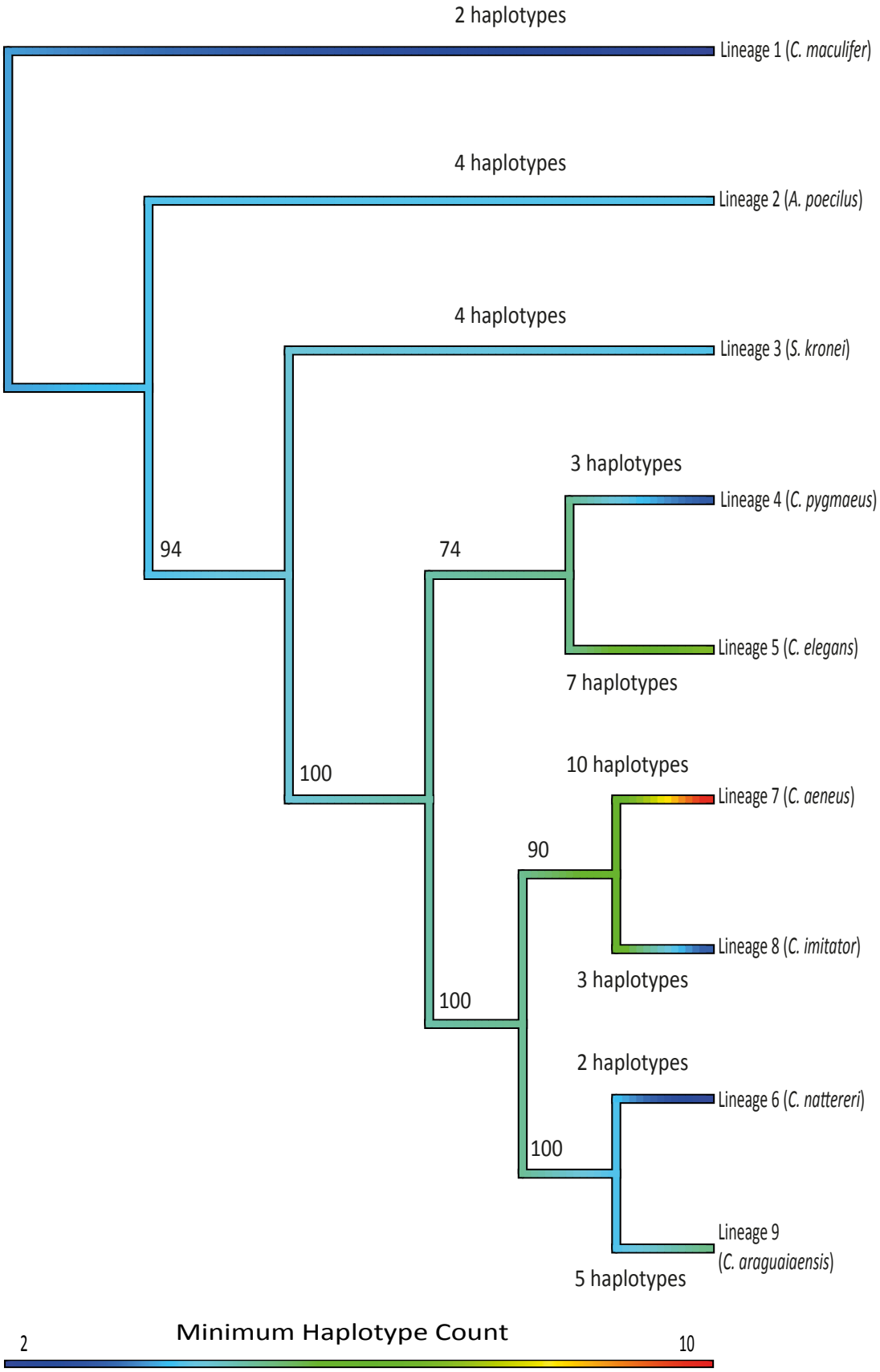


Figure 2.4: Topology recovered from the phylogenetic analysis of TLR2 genes in nine Corydoras lineages, rooted to C. maculifer (lineage 1). Trees were built in IQ-TREE utilising K2P+G4 model. Figures at nodes represent bootstrap support. Tree is coloured according to number of haplotypes identified.

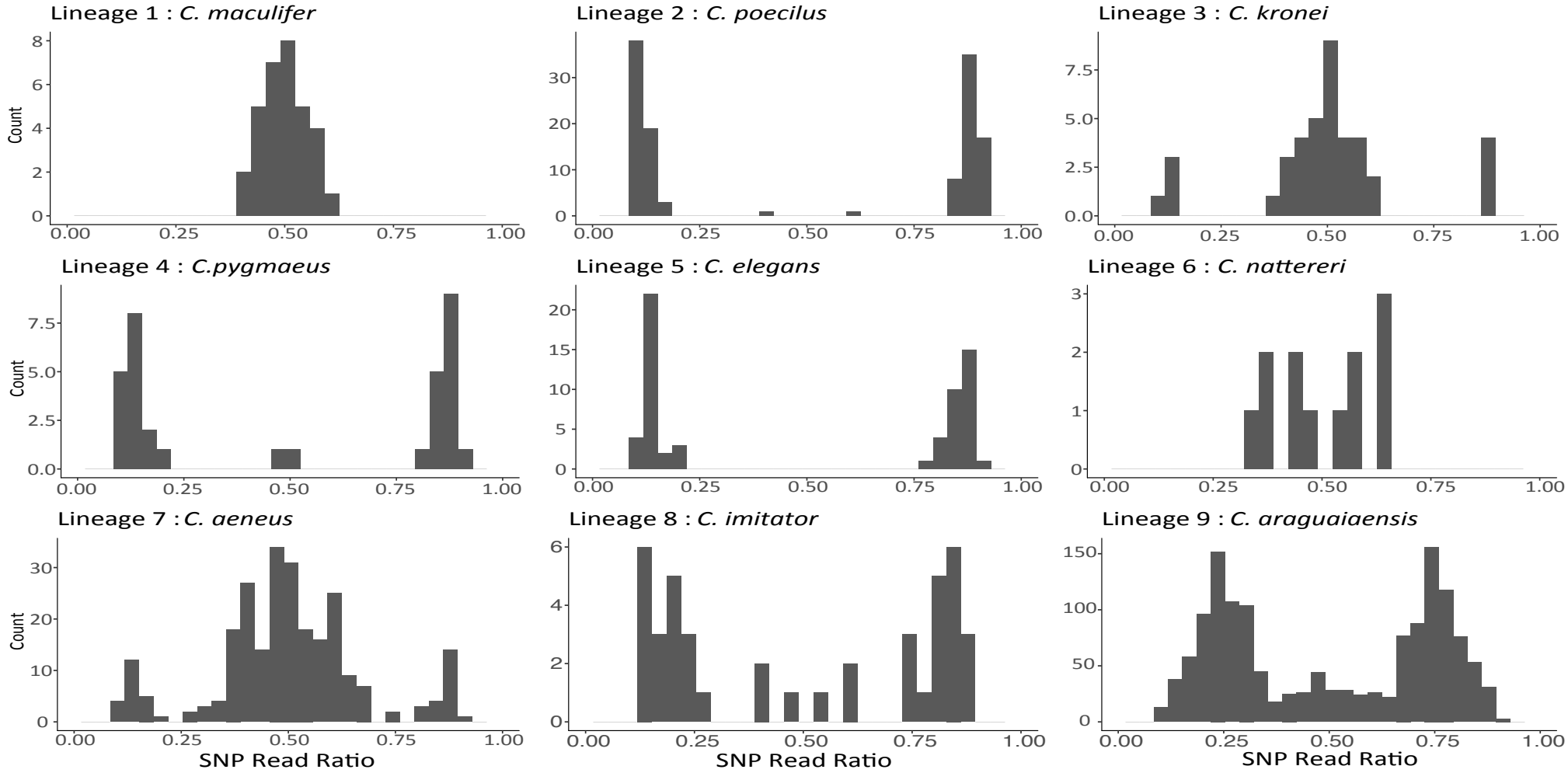


Figure 2.5: SNP read ratios within TLR2 in single individuals from lineages 2 to 8 and averaged across TLR1 and TLR2 in populations of lineage 1 (C. maculifer (n=17)) and lineage 9 (C. araguaiaensis (n=35)).

2.3.4 Variant distribution

Lineage specific amino acid sequences were submitted to SMART analysis in order to determine the location of different protein domains. These were then used as scaffolds to map SNP locations back to (Figure 2.6). The overall predicted protein structure of TLR2 did not vary much between lineages except in the number of leucine rich repeat (LRR) regions. SNPs were widely distributed across TLR2 in representatives from lineage 2, lineage 3, lineage 7 and lineage 9 and to a lesser extent in lineage 4 and lineage 5. In lineage 8 SNPs were broadly localised to the LRR C-terminal region and Toll interleukin receptor (TIR) region and in lineage 6 SNPs were only found in the TIR region. The single SNP found in lineage 1 was located in an unspecified region of TLR2.

2.3.5 Variant sharing between the nine lineages

Due to the complications associates with duplication events (WGD or tandem) and the short fragment range of NextSeq sequencing it was not possible to phase individual haplotypes for TLR2 and look for shared haplotypes across the lineages.

Shared TLR2 SNPs were plotted according to the lineages they were found in and their position along the TLR2 gene. A subset of SNPs shared between lineage 2, and lineage 7 across nucleotide positions 500-1000 suggests the possibility of a haplotype shared between these two lineages (Figure 2.7). SNPs were also frequently shared between lineages 7, 8 and 9 (Figure 2.7).

Shared SNP sites were also counted across the TLR2 sequence data set and the RAD sequencing data set (from Marburger et al. 2018). Shared SNPs were plotted across the nine lineages as a heat map (see Figure 2.8 A and B), which showed a trend towards higher SNP sharing in higher lineages. In addition, lineage 2 and 7 shared a disproportionately high number of SNPs given their relative phylogenetic positions. Variant sharing across RAD data also showed a general increase in shared SNPs among higher lineages that diverged more recently. However, lineage 5 showed a surprisingly high number of shared SNPs given its placement phylogenetically.

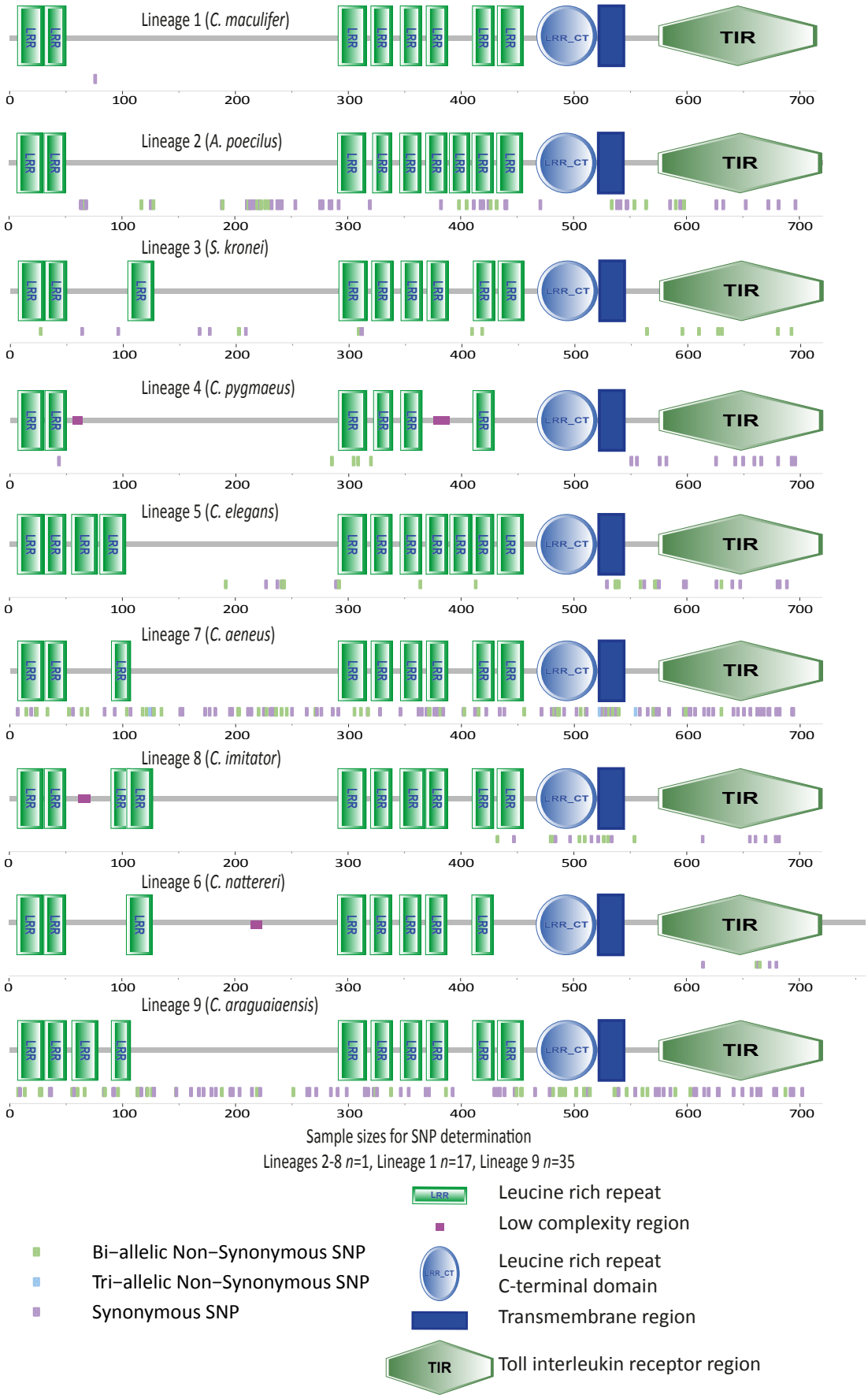


Figure 2.6: TLR2 domains inferred from SMART analysis and SNPs identified per species mapped according to amino acid position from representatives across the nine Corydoradinae lineages.

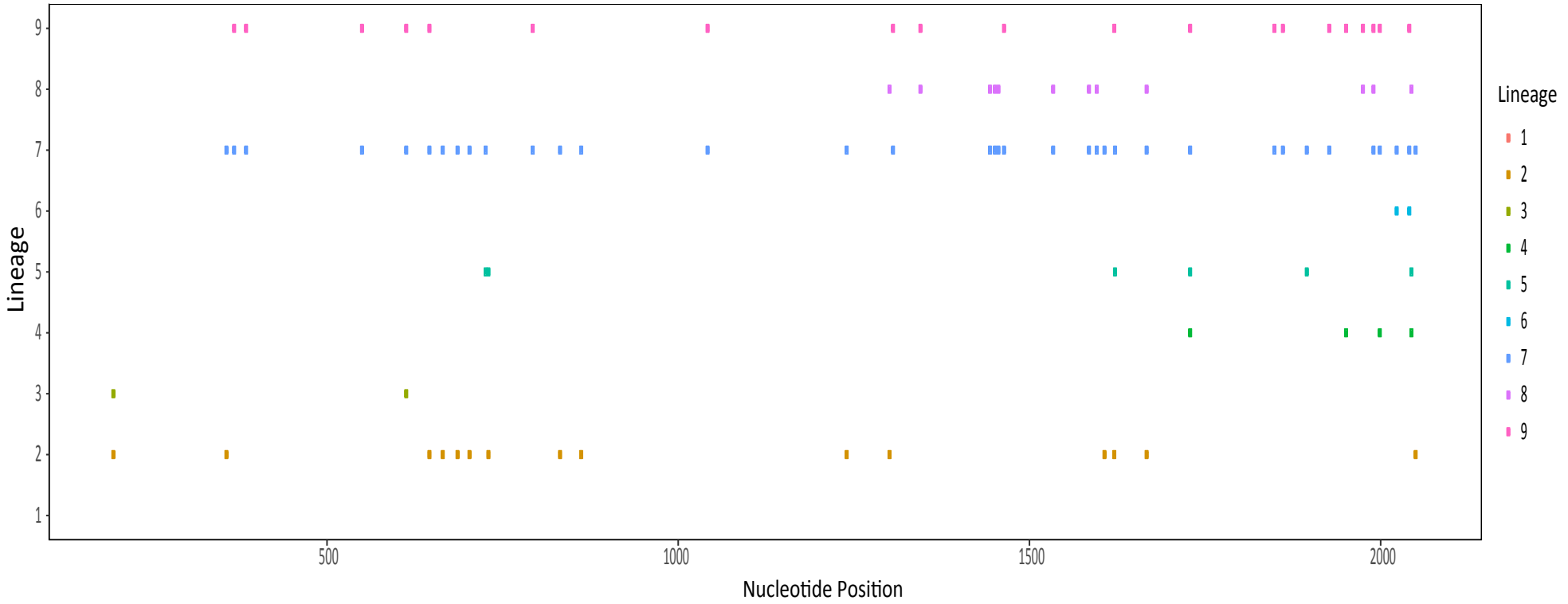


Figure 2.7: SNPs shared in at least two Corydoradinae lineages mapped according to their nucleotide position along TLR2. Lineages 2-8 were represented by single individuals, lineage 1 was represented by 17 individuals and lineage 9 was represented by 35 individuals

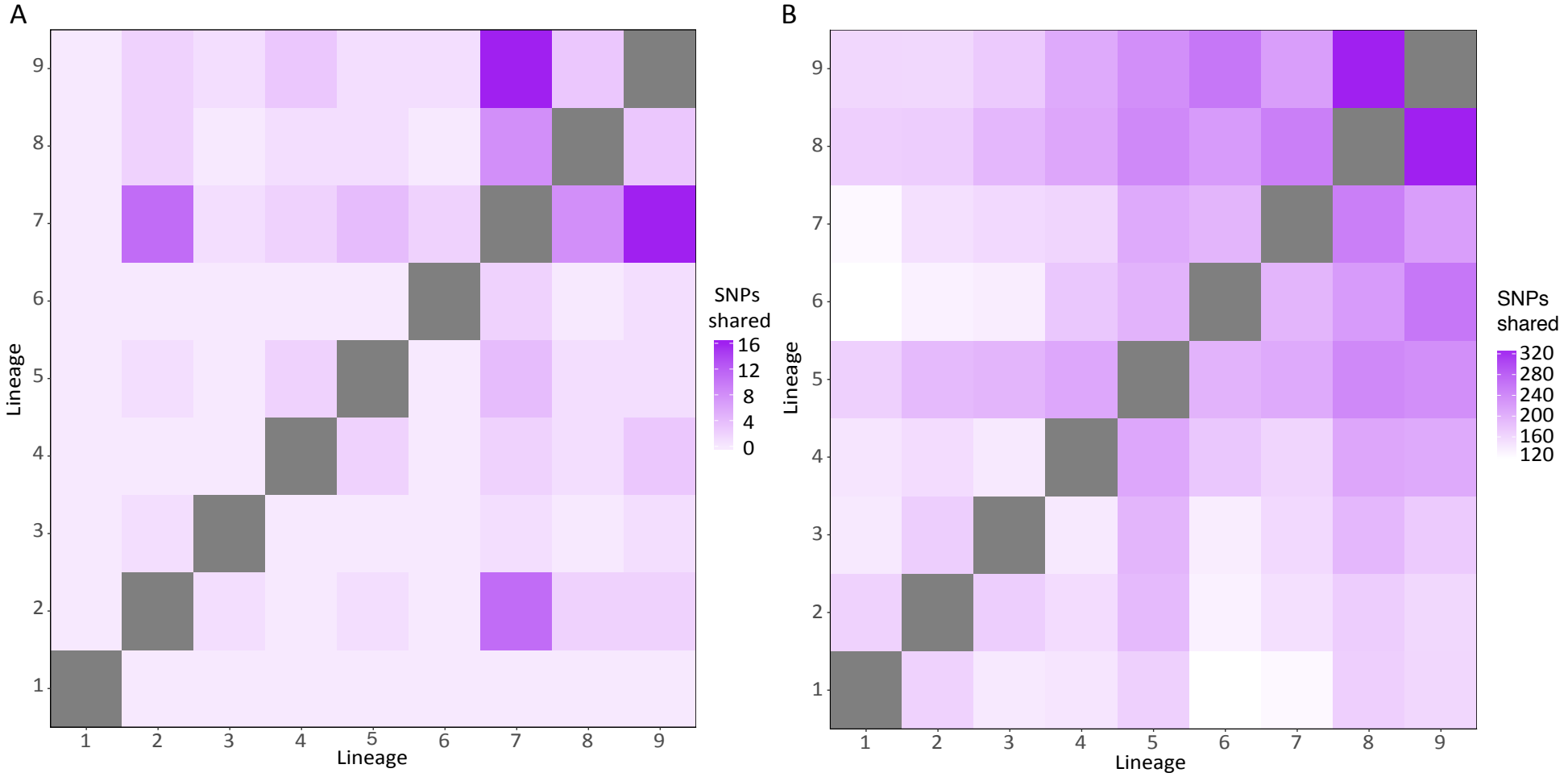


Figure 2.8: A) Number of SNPs shared across the nine Corydoradinae lineages in TLR2. Lineages 2-8 were represented by single individuals, lineage 1 was represented by 17 individuals and lineage 9 was represented by 35 individuals B) Number of SNPs shared across RAD sequence data from all nine lineages, each lineage represented by two individuals.

2.4 Discussion

In this chapter, genetic diversity in a single TLR gene (TLR2) was assessed across species of Corydoradinae catfishes from across the nine lineages identified in this subfamily. The aim of this was to firstly characterise TLR2 variation in this subfamily and then compare diversity and haplotype retention in this immune gene to similar metrics derived genome wide loci (RAD sequence data). Haplotype retention across RAD sequence data was found to be lowest in lineage 1 (as expected in a diploid lineage) and markedly highest in lineage 9 with variable levels of retention in the remaining seven lineages. When assessed across RAD sequence data, SNP ratios per contig were observed to generally increase in the higher lineages, with an exception of lineage 6 which showed SNP ratios similar to lineage 9. SNP sharing was also evident across the RAD data, with higher lineages (which have diversified more recently) generally sharing more SNPs.

SNP and haplotype estimates from TLR2 sequencing data were hampered by sample size (i.e. the fact that we only had one sample in 7 of the lineages) but showed markedly different patterns from those observed in the RAD data. SNPs were notably highest in lineage 7 (*C. aeneus*) and lineage 2 (*A. poecilius*) and lowest in lineage 1 (*C. maculifer*) and lineage 6 *(C. nattereri*). Haplotype estimates suggested that all lineages, except lineage 1 (*C. maculifer*) and lineage 6 (*C. nattereri*), had retained more than two haplotypes of TLR2. Lineage 9 (*C. araguaiaensis*) showed evidence of carrying up to four haplotypes while the remaining lineages showed evidence of carrying 6 to 10 TLR2 haplotypes. The distributions of SNPs across the nine lineages were broad and, given that they were calculated in most cases from single individuals, no more specific associations with specific domains of the TLR could be ascertained. When examining SNP sharing between different lineages across TLR2 the expected tendency for higher lineages to share more SNPs was observed. However, lineage 2 and lineage 7 were notable outliers, being phylogenetically distant and yet sharing more SNPs with each other than with more closely related lineages. A subset of SNPs shared between lineage 2, and lineage 7 between positions 500bp-1000bp suggests the possibility of a haplotype shared between these two lineages, but without further read phasing it is impossible to comment further. SNPs were also frequently shared between lineages 7, 8 and 9 which again might indicate shared haplotypes but which could only be confirmed if read phasing was possible (Figure 2.7).

The TLR2 specific analyses are based on single individuals in lineages 2 to 8, with such a small sample size it is possible that the data are not representative of their respective populations. However, they do raise a number of interesting points to discuss. The first main discussion point relates to the high levels of SNPs observed in lineage 2 (*A. poecilius*) and 7 (*C.*

aeneus) - and the high proportion of these shared between these lineages (potentially indicating shared haplotypes). These patterns could be indicative of introgression, convergence or incomplete lineage sorting between lineages 1 and 7 (Těšický and Vinkler, 2015). Evidence from phylogenetic topologies does not support introgression between the two lineages. However, because the trees were based on consensus sequences, rather than on individually phased haplotypes, it is possible that any signature of introgression might have been lost. Convergence - i.e. the independent evolution of similar genetic features (e.g. SNPs) - is a possibility; this is unlikely given the number of SNPs shared between several of the lineages. Incomplete lineage sorting under balancing selection conditions could have allowed shared SNPs/haplotypes to persist in relatively distantly related lineages, so this may be a possible mechanism for the retention of shared diversity (Těšický and Vinkler, 2015). The *Corydoras* live in mixed species communities with representatives of at least two lineages (Alexandrou *et al.*, 2011). It is therefore possible for relatively evolutionarily distant species to still share similar environmental and pathogenic influences which may develop into similar selection pressures. It is also possible that some of the species that share less immediate environmental locations may still share similar pathogen-based selection pressures and therefore be subject to balancing selection. Unfortunately, sequences could not be phased into individual haplotypes so analysis of selection and full haplotype comparison between the lineages could not be completed. This along with further sequencing data over more individuals and TLR genes would provide a more complete picture and may provide further evidence to support the possibilities that these data hint at.

The second main discussion point from the results is the high level of TLR2 haplotype retention across all lineages excepting lineage 1 (*C. maculifer*) and lineage 6 (*C. nattereri*). This haplotype retention, most notably within lineage 7 (*C. aeneus*), is higher than would be expected to have arisen and maintained from the suspected WGD events in the evolutionary history of the Corydoradinae (Marburger *et al.*, 2018). The TLR family of genes has a history of tandem duplications across the animal kingdom (Hughes and Piontkivska, 2008; Temperley *et al.*, 2008). The results of this chapter might be seen as evidence for further duplications within some of the Corydoradinae lineages, or of the differential loss across lineages of tandem replicates that occurred early in the diversification of the Corydoradinae. However, given the low sample size here further sequence analyses, across a greater range of individuals and TLR genes, would be required to make more substantive claims.

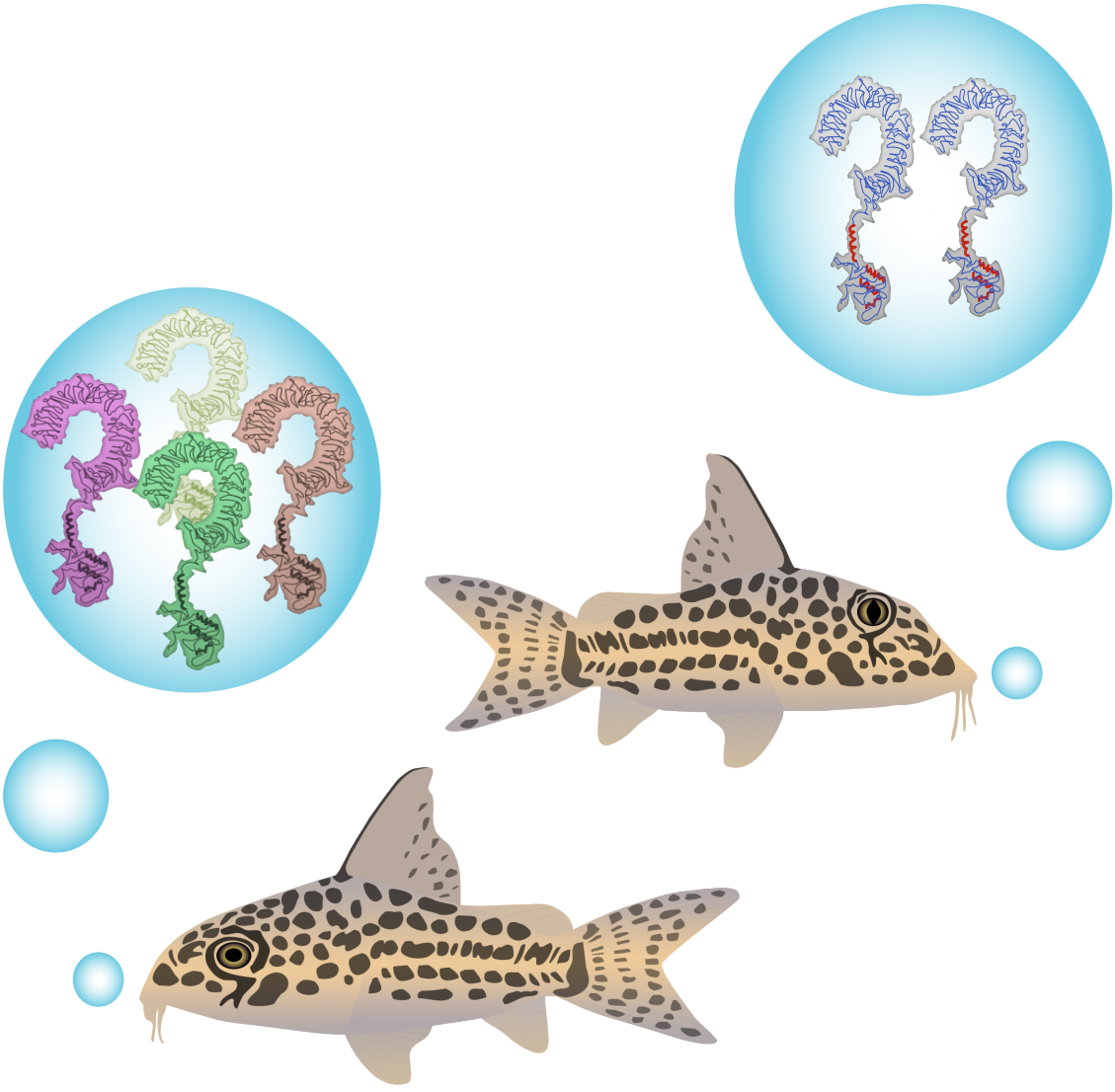
2.4.1 Conclusion

In summary this chapter has aimed to characterise TLR2, explore TLR2 specific SNP diversity, haplotype retention and variant sharing between the nine Corydoradinae lineages and

compare these metrics to representative portions of the relevant genomes via RAD sequencing data. It is limited by the number of individuals sequenced, a lack of haplotype phasing and its focus on a single TLR gene. However, the unexpectedly high numbers of retained haplotypes in some of the lineages does provide potential evidence for lineage specific tandem duplication events within the evolutionary history of the TLR2 gene, and highlights potential incomplete lineage sorting or introgression across the nine Corydoradinae lineages.

Chapter 3:

Toll-like Receptor variation within diploid and polyploid *Corydoras* catfishes



3.1 Introduction

The immune systems of animals serve as defence mechanisms against invading pathogens. These systems are divided into innate and adaptive immune pathways. Both pathways are composed of a range of specialised cells and proteins adapted for host defence (Takeda and Akira, 2005). A pivotal part of the function of immune proteins is the recognition of foreign antigens and the successful mounting of an immune response (Takeda and Akira, 2005). As a result, diversity in and among immune genes and the proteins they encode may be favourable for the recognition of a greater range of potential pathogens; this is well documented in several immune gene families including the extremely polymorphic Major Histocompatibility Complex (MHC) loci (Zinkernagal and Doherty, 1974; Hill, 1999). There are a number of theoretical mechanisms that could assist with the expansion and maintenance of immune gene diversity (Spurgin *et al.*, 2011; King, Seppälä and Neiman, 2012)

Whole genome duplication (WGD) events, where all the genetic material in a given organisms is duplicated one or more times, represent a potential mechanisms for increasing genetic diversity and facilitating adaptive radiation, although robust empirical investigations are lacking (Mable, Alexandrou and Taylor, 2011). Nevertheless, WGD events have been associated with the rise of highly species rich groups, such as flowering plants and teleost fishes (Masterson, 1994; Peer, Maere and Meyer, 2009; Pasquier *et al.*, 2016). Such WGD events are frequently followed by genome rearrangements and re-diploidisation via gene fractionation wherein many duplicated gene copies are functionally silenced or deleted (Berthelot *et al.*, 2014). However, in the case of immune genes, there may be an intrinsic advantage to maintaining additional gene copies, as they may benefit host defence via expanding pathogen recognition mechanisms through increased diversity or increasing dosage during transcription (King, Seppälä and Neiman, 2012).

The relationship between polyploidy and functional genetic diversity is not straightforward and a number of theoretical outcomes have been suggested. Gene copies duplicated through WGDs (ohnologues) may be more subject to genetic drift and mutation than alleles in diploid organisms, and this may increase the probability of an individual carrying a beneficial mutation (Otto and Whitton, 2000). However, if selection is weakened in polyploid lineages, compared with diploid lineages, the combination of drift and selection may result in non-adaptive radiations (Paquin and Adams, 1983; Gorelick and Olson, 2013). Weakened selection is theoretically expected because of the greater number of alleles per locus (Otto and Whitton, 2000). This may slow the spread of beneficial alleles through a population and also mask, and therefore preserve, deleterious mutations (Otto and Whitton, 2000). Finally, and in direct contrast to the previous mechanisms, the presence of additional gene copies may allow

the basic function of the gene to be maintained while duplicated copies are effectively freed from selection. These duplicated copies may then diversify in novel directions (genetic innovation), increasing adaptive potential (Otto and Whitton, 2000).

Additional factors are relevant when considering the effects of polyploidy on population wide immunogenetics. Firstly, polyploid individuals are more likely to be heterozygous at a given locus (King, Seppälä and Neiman, 2012). Heterozygote advantage is a condition whereby the immune system of heterozygous individuals can detect a broader array of pathogen peptides and therefore recognise and respond to a greater range of pathogens or peptides from the same pathogen and so eliciting a greater immune response than non-heterozygous individuals. If heterozygote advantage comes into play a role in immunity, then polyploid individuals may have higher fitness than diploids (Doherty and Zinkernagel, 1975; King, Seppälä and Neiman, 2012). Secondly, polyploids are more likely to carry rare alleles. Pathogens may develop defence or evasion mechanisms to commonly occurring immune variants, giving advantage to organisms carrying rare versions of those genes (Otto and Whitton, 2000; King, Seppälä and Neiman, 2012) – a process termed 'negative frequency dependence' (Slade and McCallum, 1992). Other mechanisms, such as gene conversion (which may act to generate or obliterate gene diversity) (Spurgin *et al.*, 2011), alloploid associated independent gene evolution (where duplicated genes evolve independently of each other but match the rates of their progenitor ancestors) (Cronn, Small and Wendel, 1999), and rapid genome change in early polyploidisation (Song *et al.*, 1995) may also have impacts on immunogenetic diversity. However, the conditions under which these mechanisms act appear to be case specific.

In animals one of the primary mechanisms of the innate immune system are the pathogen recognition receptors (PRRs) (Rajendran *et al.*, 2012). PRRs are germ line encoded, have broad specificities and are adapted to recognise evolutionary conserved pathogen-associated molecular patterns (PAMPs) (Alvarez-Pellitero, 2008). These PAMPs tend to be of large importance for pathogen virility or survival and as a result are resistant to evolutionary alterations (Janeway 1989 as sited in Medvedev 2013). In fish species, these PRR types include toll-like receptors (TLRs), nucleotide binding oligomerization domain (NOD)-like receptors (NLRs) and retinoic acid inducible gene I (RIG-I)- like helicases (RLHs) (Aoki and Hirono, 2006; Chang *et al.*, 2011; Rajendran *et al.*, 2012).

The TLRs represent a group of type I transmembrane proteins which recognise extracellular PAMPs and initiate innate immune mechanisms when activated (Zhao *et al.*, 2013). They share a common structure, which includes: an N terminal ectodomain, multiple leucine rich repeats (LRRs), a C terminal cytoplasmic tail, a transmembrane region and a toll

interleukin receptor (TIR) signalling domain (Medvedev, 2013). TLRs are thought to play a direct role in activation of the innate inflammatory response, and may influence adaptive immune responses through regulation of antigen presentation on dendritic cells, and through direct effects on T and B lymphocytes (Salaun, Romero and Lebecque, 2007). They may also induce apoptosis (Salaun et al., 2007). Subfamilies of the TLRs are broadly associated with different pathogenic groups. In fish TLR1, TLR2, TLR4, TLR5 and TLR9 have been associated with bacterial infection while TLR3, TLR22, TLR7 and TLR8 appear to be associated with viral infections (Fink *et al.*, 2016). These specialisations are relatively broad and do not rule out other functions. For example TLR1, TLR2, TLR9, TLR19, TLR21 and TLR25 have also been associated with responses to *Ichthyophthirius multifiliis* (a eukaryotic parasite) infection in channel catfish (Zhao *et al.*, 2013).

The genes that encode TLRs are highly polymorphic (Netea, Wijmenga and O'Neill, 2012). A number of genome wide association studies (GWAS) have highlighted potential associations between specific polymorphisms in the TLRs and susceptibilities to disease in humans (Skevaki *et al.*, 2015), although these results have been questioned (Netea, Wijmenga and O'Neill, 2012). Phylogenetic evidence suggests that some of the major functional specialisations of the TLRs arose following ancient gene duplication events, prior to the divergence of protostomes and deuterostomes, and these traits have been preserved through later gene duplication events (Hughes and Piontkivska, 2008). For example, specialised recognition of bacterial lipoproteins appears to have evolved in an ancestor of the TLR1 subfamily and persists as a recognised characteristic in TLRs within this subfamily (Hughes and Piontkivska, 2008). A loss of the major histocompatibility complex class II (MHCII) following ancient genome expansion in Gadiformes lineages is thought to be an influential factor in the development of new innovations in the TLR gene family (Solbaken et al., 2017). Additionally, while multiple copies of some TLRs have been observed in a number of teleost species, additional copies of TLR1, TLR2, TLR2 and TLR5 were not observed (Solbakken *et al.*, 2017).

In catfishes, the TLR gene family (including TLR1, TLR2, TLR9, TLR19, TLR21 and TLR25) has been shown to have change expression profiles in channel catfish *Ictalurus punctatus,* infected with the parasite *I. multifiliis* and may, therefore have a function in parasite associated responses in *Corydoras* catfishes (Zhao *et al.*, 2013).

3.1.1 Aims and objectives

Here we investigate the relationships between genome size expansion and differences in functional TLR genes in two coexisting species of *Corydoras* catfishes that exhibit markedly different genome sizes. *Corydoras maculifer* (lineage 1, diploid) has a C value of 0.5pg and

Corydoras araguaiaensis (lineage 9, putative tetraploid) has a genome size of 4.2pg (Marburger, 2015). Both species are sympatric in the Rio Araguaia catchment in Matto Grosso, Brazil. Thus, their environment is the same and both species should be exposed to similar pathogenic communities, allowing a direct comparison of genome size and TLR diversity.

We use Next Generation Sequencing (NGS) on an Illumina NextSeq platform to characterise genetic diversity in the complete coding sequence (CDS) of two TLRs (1 and 2) across population samples of *C. maculifer* and *C. araguaiaensis*. We first characterise the structure of these TLRs in the two species and then measure the genetic diversity of the genes. Based on the predicted increase in genetic diversity after WGD, we predict that the putative tetraploid, *C. araguaiaensis,* will have greater immune gene diversity than the diploid *C. maculifer*.

3.2 Methods

3.2.1 Sampling and DNA extraction

Wild living individuals of *C. maculifer* (n=17) and *C. araguaiaensis* (n=36) were collected from the same location in the Araguaia region in Brazil by MIT, CO (2012 and 2015) and EB (2015), euthanised by anaesthetic overdose and stored individually in 100% ethanol. DNA was extracted from fin clip tissue using the salt extraction protocol after Sunnucks & Hales (1996) and Aljanabi & Martinez (1997).

3.2.2 PCR amplification and library preparation

Two TLRs (1 and 2) were PCR amplified in 17 *C. maculifer* samples and 36 *C. araguaiaensis* samples according to the methodologies described in Chapter 2. TLR amplicons from the same individual were pooled in equimolar concentrations and submitted to the library preparation protocols described in Chapter 2.

3.2.3 Data processing and analysis

All sequencing data was quality checked, de-multiplexed by ligation index and then by PCR barcode, and cleaned to remove traces of adaptor contamination as described in Chapter 2. Reads from a single individual from each species were mapped to TLR1 and TLR2 sequences from the *C. maculifer* genome (MIT unpublished) data using BWA-mem (version 0.7.12, Li & Durbin 2009) and consensus sequences produced using Geneious-9.0.5. These consensus sequences were edited manually to replace ambiguous bases with non-ambiguous base codes, and were used as species-specific references for subsequent mapping of raw reads and SNP calling. Reads from all individuals of each species were mapped to these single individual derived consensus sequences using BWA-mem prior to SNP calling across populations with FreeBayes (version 1.1.0, Garrison & Marth 2012). Variants were called initially using FreeBayes and then filtered using the FreeBayes CSV filter for quality scores of greater the 20, a mean mapping score for alternative alleles of greater then 40, and reference and alternative allele observation counts of greater than 5. The resulting data were filtered further in MS excel. In order to be included in downstream analysis a SNP had to have a read depth of greater then 10% of the overall read depth at that site. This was to remove artefacts of PCR or sequencing error. All data points were plotted using ggplot2 (version 2.2.1) within R studio (R version 3.4.1). Where applicable Wilcoxon statistical tests were used to determine significance between species.

Phylogenetic trees for each TLR locus were built using IQ-TREE (version 1.5.5, Nguyen et al. 2015; Hoang et al. 2018), which builds maximum likelihood trees based on the best model fit identified by jModelTest using the Bayesian information criterion (BIC). Trees were built using 1000 ultrafast bootstrap replicates. Trees were visualised using FigTree (version 1.4.3) and included *I. punctatus* TLR sequences as out-groups.

Observed and expected heterozygosity was calculated using the traditional Hardy-Weinberg equilibrium (Equation 1) for *C. maculifer.* An adapted version of the same equation modified for tetraploids (Equation 2; Frankham et al. 2010) for *C. araguaiaensis*. For the purposes of this equation we assume that *C. araguaiaensis* is an autopolyploid tetraploid and SNP dosage was calculated as an estimate according to this premise. Observed and expected heterozygosity was calculated for each SNP site and values were averaged across each gene.

*Equation 1: diploid expected genotype frequencies (Hardy-Weinberg)**

$$p^2 + 2pq + q^2 = 1$$

*Equation 2: tetraploid expected genotype frequencies**

$$p^4 + (4p^3q + 6p^2q^2 + 4pq^3) + q^4 = 1$$

**Where p and q refer to expected allele frequencies*

Average counts of synonymous and non-synonymous SNPs were calculated for TLR1 and TLR2 in *C. araguaiaensis* and ratios of synonymous to non-synonymous substitution were estimated. Dn/Ds ratios could not be calculated because individual haplotypes could not be phased (see below).

Complications inherent in polyploid sequencing data and the uncertainty regarding the allopolyploid or autopolyploid status of the study species means that reliably phasing short reads into individual full-length haplotypes was not possible with these data. In order to investigate patterns in SNP presence among individuals of the *C. araguaiaensis* population, SNP profiles (synonymous and non-synonymous combined) were plotted for both TLR genes, and just non-synonymous SNP substitutions to show only functional variation. These profiles compare location and prevalence (homozygote or heterozygote) of each SNP for each individual from within the *C. araguaiaensis* population. *C. maculifer* only exhibited a maximum of one SNP per gene so this step of the analysis was considered redundant for this species.

Haplotype estimates (defined as: the minimum number of unique alleles or haplotypes found within a single individual) were initially calculated for both species. QualitySNPng (Nijveen *et al.*, 2013) estimates SNP counts, but also the number of haplotypes across short distances of overlapping reads (short range phasing) for each individual. However, this program has a tendency to overestimate haplotype numbers (based on personal observation),

so all outputs were manually validated. A second method for ascertaining haplotype number across populations was undertaken by calculating the read depth ratio of each allele. Biallelic SNPs present in a diploid should have a ratio of 0.5 as half of reads (those from one haploid genome copy) have the reference base and half will have the alternative base (i.e. on the other haploid genome copy). In a triploid, peaks in the number of SNP reads should occur at frequencies of 0.33 and 0.66 and in tetraploids, SNP ratios of 0.25, 0.5 and 0.75 would be expected (Marburger, 2015).

The frequencies of alternative bases for each species were plotted along the length of both TLR genes to determine any spatial pattern to SNP location, e.g. do they appear at greater frequency within specific TLR functional regions. Variation was plotted using two methods; SNP frequencies were first calculated by dividing the occurrence of an alternative base by the total number of individuals of each species regardless of ploidy status or homo/heterozygosity. The second method assumed diploidy in C*.* *maculifer* and tetraploidy in C*.* *araguaiaensis*. The proportion of the read depth each alternative base occupied was calculated, assigned to a bin (1-2 in the diploid or 1-4 in the tetraploid), summed across populations, and divided by the total number of expected haplotypes across the population (2X total number of diploids or 4X total number of tetraploids). Protein domains were identified across sequences for both TLRs in both species of *Corydoras* using Simple Modular Architecture Research Tool (SMART, Letunic & Bork 2018). Single consensus sequences for each gene and each species were fed through the SMART algorithm, which identified and annotated signatures of protein domains.

3.3 Results

3.3.1 Sequencing, Data Cleaning and Quality Control

Sequence data for *C. maculifer* and *C. araguaiaensis* were obtained from the same sequencing lane as the data for Chapter 2. Reads retained at each of the clean-up and mapping stages for *C. maculifer* and *C. araguaiaensis* are listed in Table 3.1 along with mean read depth following mapping.

Consensus sequences for each individual for *C. maculifer* and *C. araguaiaensis* were derived using Geneious V9 prior to alignment to TLR sequences for *I. punctatus* using MUSCLE (through Geneious V9). Consensus alignments were used for phylogenetic analysis with IQ-TREE. Figure 3.1 represents topology for TLR consensus sequences for *C. maculifer* and *C. araguaiaensis,* with *I. punctatus* TLRs as an out-group for rooting purposes. The tree topologies show a clear distinction between the two species, indicating that contamination between species in the sequence data is not an issue. This distinction between species also suggests that there is no signal of hybridisation between the two coexisting *Corydoras* species and that there is no evidence of trans-species polymorphism in TLR1 or TLR2 as you would expect some overlap between species if this were the case.

Table 3.1: Read retrieval each analysis stage following the single ended sequencing run in C. maculifer and C. araguaiaensis.

Species	**Lineage**[1]	**Number of loci sequenced**	**De-multiplexed reads**	**Mapped reads**	**Mean depth**
C. maculifer (n=17)	1	2	6689 (SD 11096)	2004 (SD 910)	62.3 (SD 28)
C. araguaiaensis (n=36)	9	2	3951 (SD 1223)	3337 (SD 732)	103.5 (SD 23)

1: Marburger, 2015

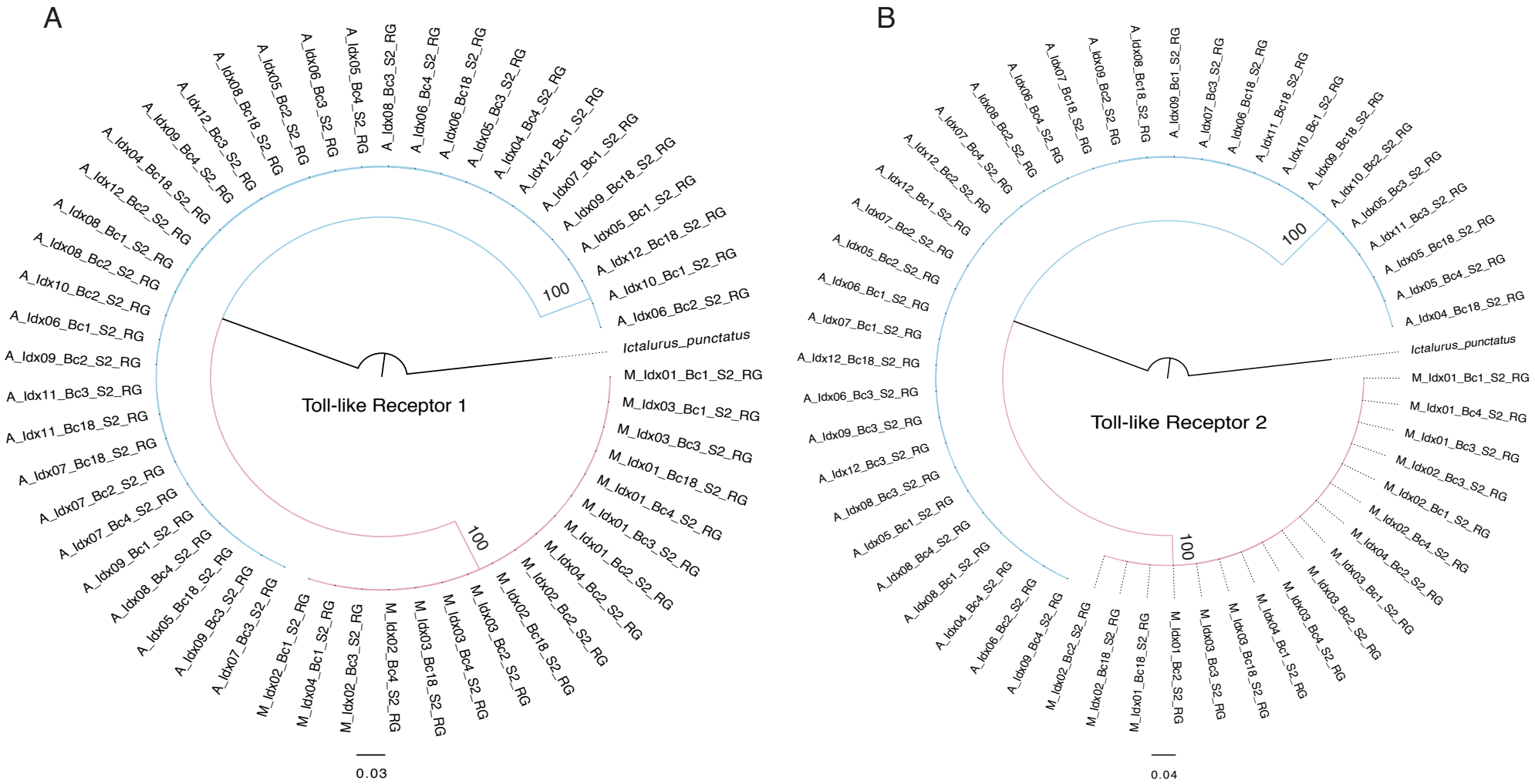


Figure 3.1: Topology recovered from the phylogenetic analysis of TLR1 (A) and TLR2 (B) genes in C. maculifer (tip label M) and C. araguaiaensis (tip label A) rooted with Ictalurus punctatus. Trees were built in IQ-TREE utilising HKY model for TLR1 (a) and TPM3+G4 for TLR2 (b); figures at nodes represent bootstrap support.

3.3.2 Variant calling

Variant (SNP) counts for both TLR genes between within populations of *C. maculifer* and *C. araguaiaensis* are presented in Figures 3.2. The SNP counts were significantly higher for both TLR1 and TLR2 in *C. araguaiaensis* than in *C. maculifer*. In both TLR1 and TLR2, *C. maculifer* had a maximum of one SNP per individual, while *C. araguaiaensis* had a total of 42 SNPs in TLR1 (with as many as 10 SNPs being found within an individual) and 114 SNPs in TLR2 (with a maximum of 29 SNPs being identified in an individual). This pattern held when SNPs were segregated into synonymous and non-synonymous substitution types. TLR2 had a significantly greater number of SNPs than TLR1 in *C. araguaiaensis* populations (Wilcoxon Test: $W = 16.5$, $p < 0.01$) (Figure 3.2).

Observed and expected heterozygosity was calculated per SNP loci and averaged for TLR1 and TLR2 (Table 3.2). For both species observed and expected heterozygosity values were not significantly different in either gene and observed and expected values in *C. araguaiaensis* matched each other perfectly down to two decimal places.

Synonymous to non-synonymous ratios were calculated from averaged synonymous and non-synonymous SNP counts in *C. araguaiaensis.* Ratios were in favour of synonymous SNPs for both TLRs, this method of comparing ratios is not as sophisticated as dN/dS calculations so limited conclusions may be drawn (Table 3.3).

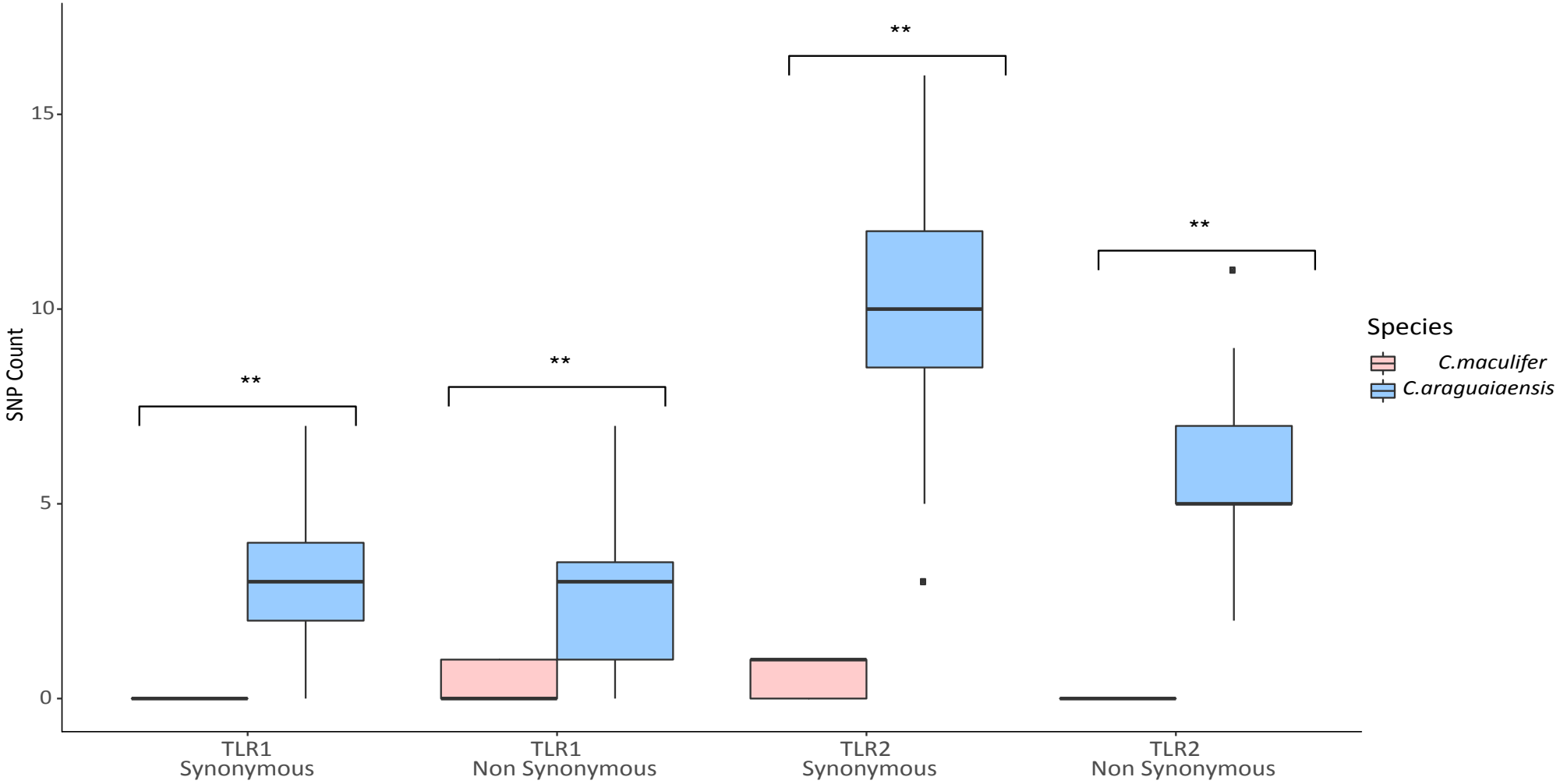


*Figure 3.2: SNP counts across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) in two Toll-like receptor genes (TLR1 and TLR2) with SNPs divided by substitution type (i.e. synonymous and non-synonymous substitutions). Significant differences between species denoted with double asterisks (** = P < 0.01).*

Table 3.2: Averaged observed and expected heterozygosity metrics for TLR1 and TLR2 in C. maculifer (diploid) and C. araguaiaensis (putative tetraploid). Calculations assume C. araguaiaensis is an autotetraploid.

C. maculifer										
	Observed Homozygote frequency		**Observed Heterozygote frequency**	**Expected frequencies**			***Chi²***		
	Reference	**Alternative**		p^2	q^2	$2pq$	X^2	***df***	***p***
TLR1	0.65	0.00	0.35	0.69	0.03	0.29	0.05	1	>0.2
TLR2	0.18	0.24	0.59	0.22	0.28	0.50	0.03	1	>0.2

C. araguaiaensis											
	Observed Homozygote frequency		**Observed Heterozygote frequency**			**Expected frequencies**			***Chi²***		
	Reference	**Alternative**	**RRRA**	**RRAA**	**RAAA**	p^4	q^4	$4p^3q + 6p^2q^2 + 4pq^3$	X^2	***df***	***p***
TLR1	0.87	0.00	0.11	0.02	0.00	0.87	0.00	0.13	7.56^{-5}	1	>0.2
TLR2	0.85	0.01	0.11	0.02	0.01	0.85	0.01	0.14	1.56^{-4}	1	>0.2

Table 3.3: Average Synonymous Non-Synonymous SNP counts per TLR in C. araguaiaensis, along with synonymous to non-synonymous SNP ratios (S:N)

	Average Synonymous SNPs count	**Average Non-Synonymous SNP count**	**S:N**
TLR1	2.94	2.48	5:4
TLR2	10	6.45	10:7

Phenotypic profiles of SNPs were constructed to investigate potential patterns in SNP presence among individuals in *C. araguaiaensis*. Profiles were first constructed to include both synonymous and non-synonymous substitutions for both TLR genes (Figure 3.3 and Figure 3.5). Profiles that included all SNPs were different for every individual examined. There may be shared haplotypes underlying the SNP profiles but each individual had a unique SNP profile for both TLR1 and TLR2. To assess how much of this diversity was functionally significant, only non-synonymous SNPs were then plotted (Figure 3.4 and Figure 3.6). In TLR1, four non-synonymous SNP profiles were shared across 11 individuals; the remaining 24 individuals had unique SNP profiles. In TLR2 all individuals exhibited unique non-synonymous SNP profiles. One of these substitutions changed the codon to a stop codon and was present in three individuals.

3.3.3 Haplotype number quantification

Analyses were conducted to infer the minimum number of possible haplotypes present for each TLR. The results from this analysis are presented in Figure 3.7. For *C. maculifer* the number of haplotypes inferred in an individual was never more then two, further supporting this species' diploid status. In contrast, *C. araguaiaensis* individuals rarely had less then two haplotypes, and as many as five haplotypes were detected in TLR2 for four individuals.

Read depth ratios were then assessed as a secondary method of ascertaining copy number. In Figure 3.8, *C. maculifer* follows the classic pattern expected of a diploid with a clear peak at 0.5, whereas *C. araguaiaensis* has two peaks at 0.25 and 0.75, the expected profile for a tetraploid.

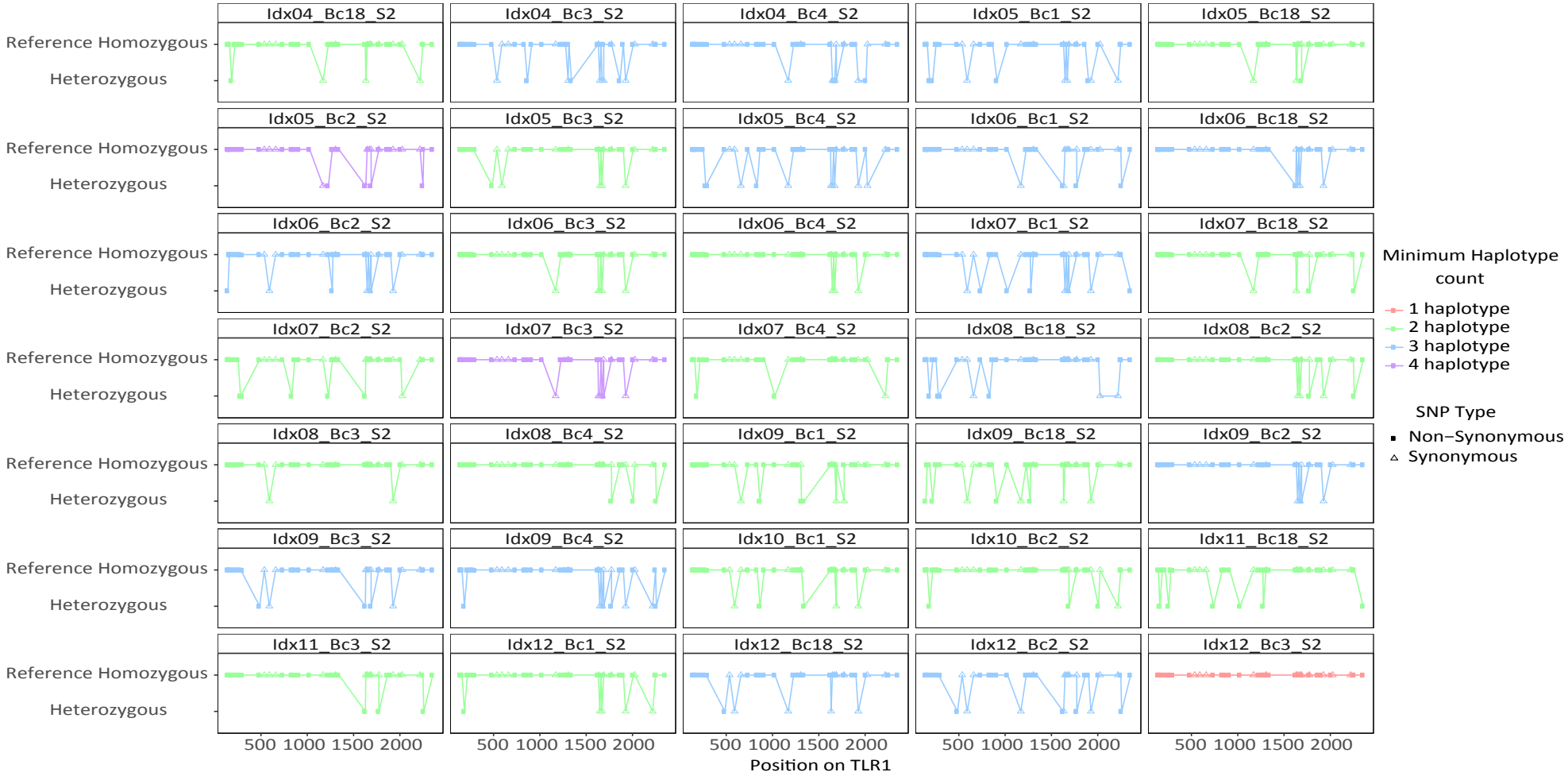


Figure 3.3: All TLR1 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP.

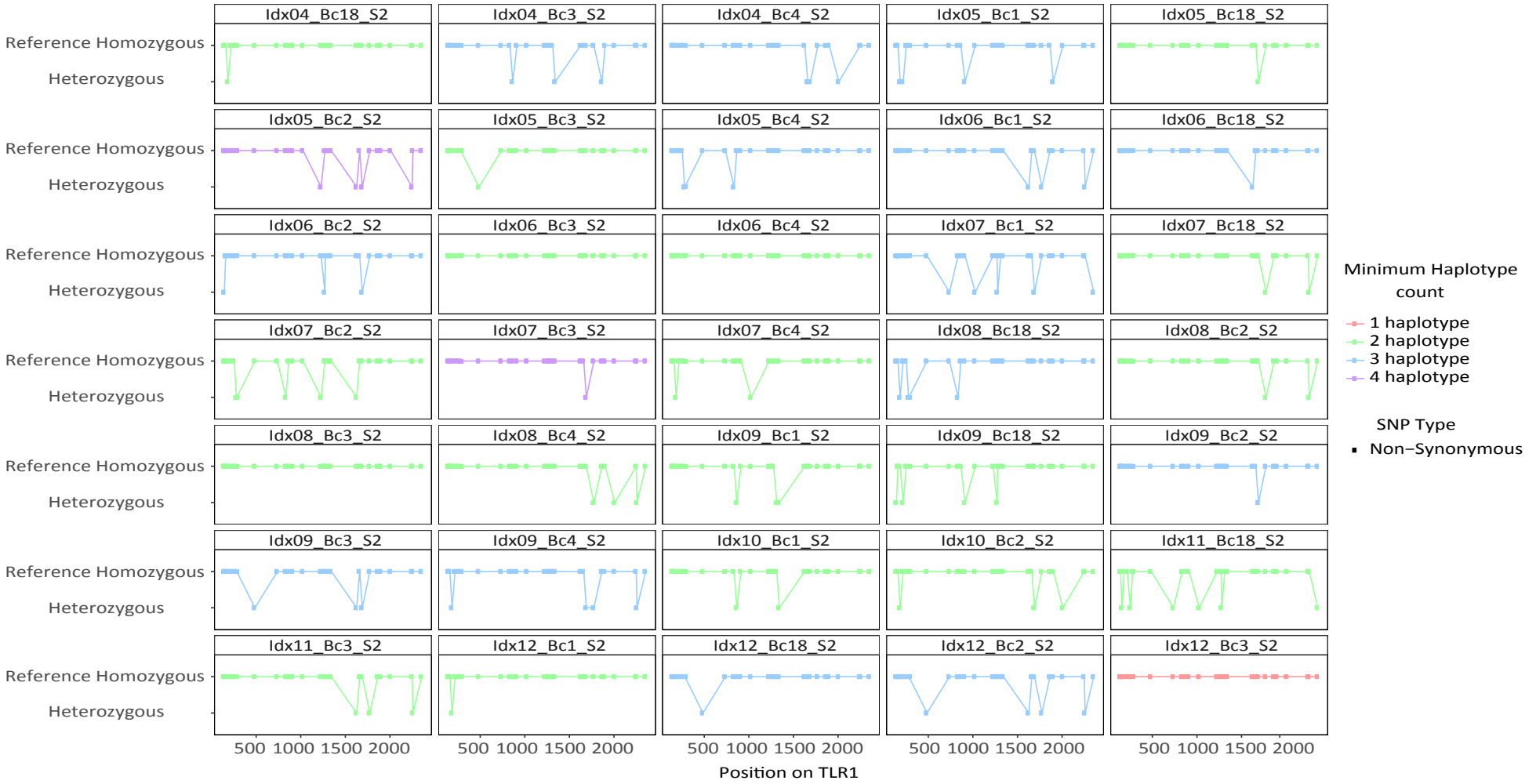


Figure 3.4: All Non-synonymous TLR1 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP.

Figure 3.5: All TLR2 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP.

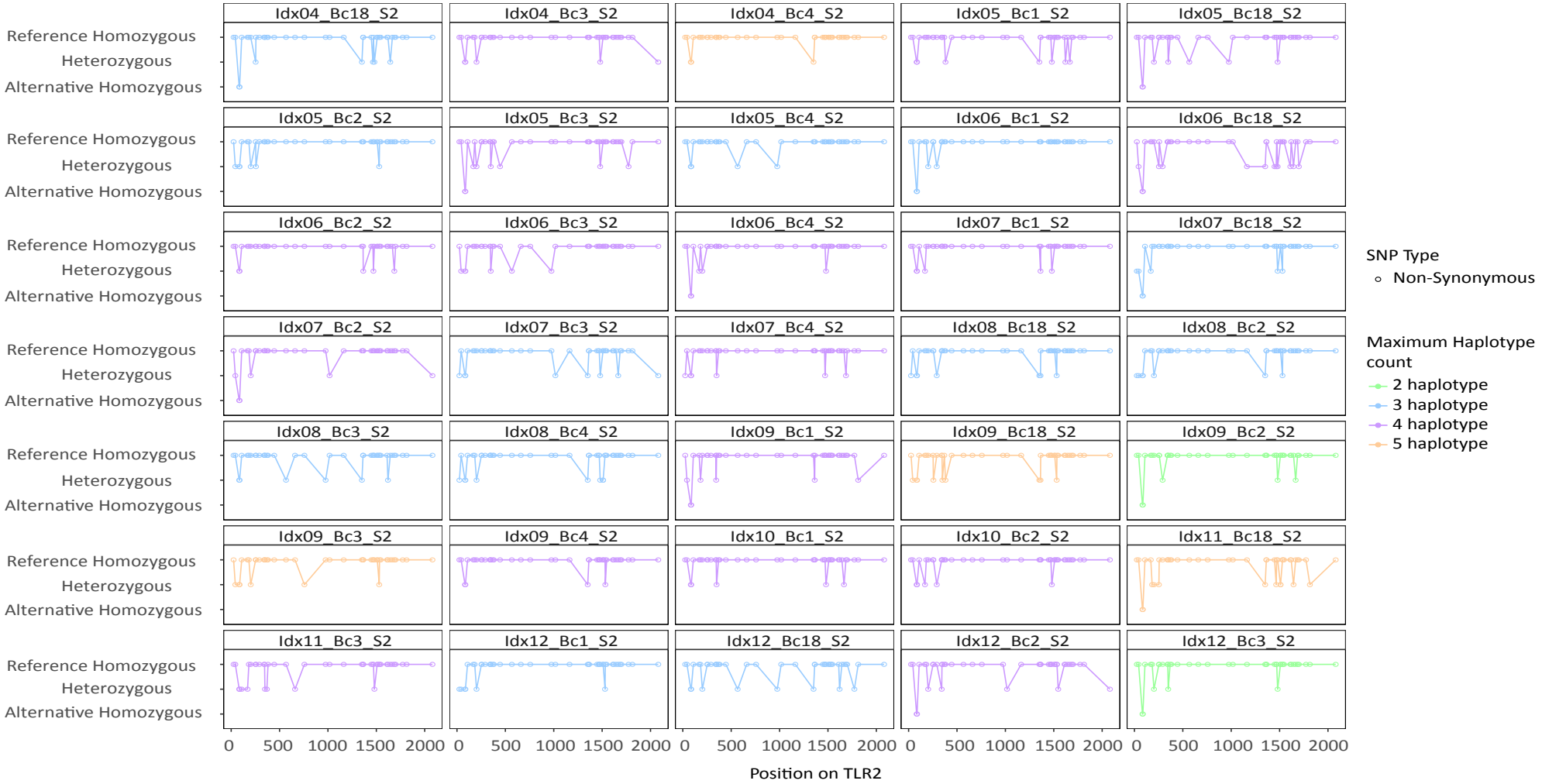


Figure 3.6: All non-synonymous TLR2 SNPs per individual C. araguaiaensis, produced as individual profiles according to SNP presence and coloured according to the number of haplotypes predicted per individual by QualitySNP.

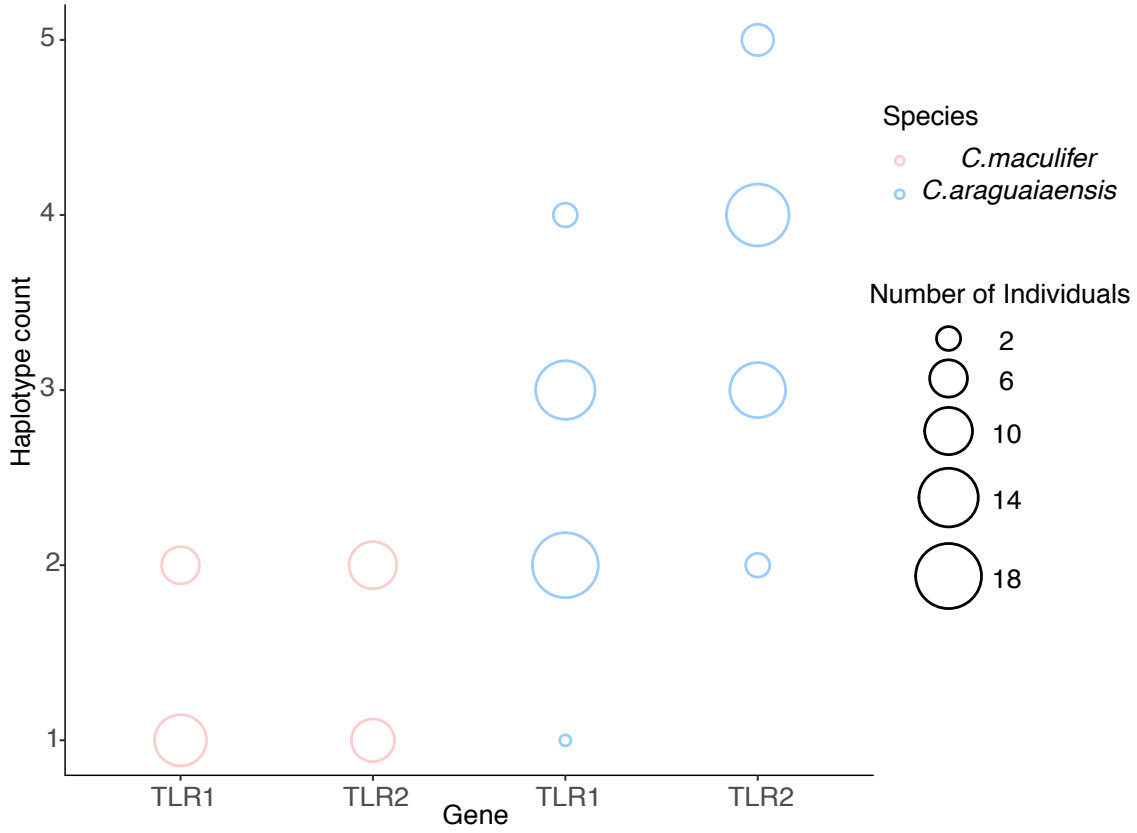


Figure 3.7: Minimum inferred haplotype counts across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) at TLR1 and TLR2 as derived by QualitySNP and verified manually.

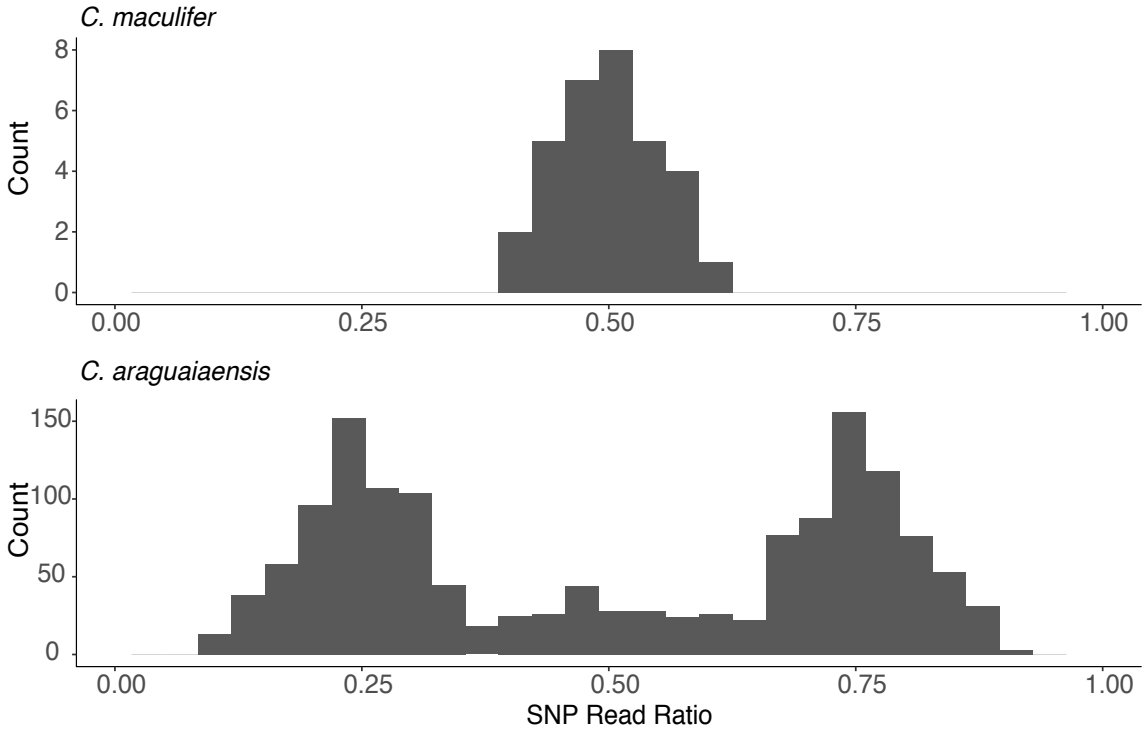


Figure 3.8: SNP read ratios averaged across populations of C. maculifer (n=17) and C. araguaiaensis (n=35) at two TLR loci (TLR1 and TLR2).

3.3.4 Toll-Like Receptor Structure

TLR1 was structurally similar between *C. maculifer* and *C. araguaiaensis* with similar placing, but differing numbers, of leucine rich repeat (LRR) domains and almost identical LRR C terminal (LRR-CT) regions and Toll Interleukin Receptor (TIR) region locations. However the transmembrane region for TLR1 in *C. maculifer* was at the start of the sequence, rather than in the more classical position situated between the LRR-CT and TIR region as observed in *C. araguaiaensis* (Figure 3.9).

The structure of TLR2 between *C. maculifer* and *C. araguaiaensis* were very similar in the placement of LRR-CT, transmembrane regions and TIR regions. *C. araguaiaensis* TLR2 had two additional LRR domains at the start of the sequence but otherwise the predicted protein structures were almost identical (Figure 3.10).

3.3.5 Variant distribution and frequency

In TLR1 in *C. maculifer* the only SNP was non-synonymous, in the LLR C terminal region, and present in *c.*35% of the population (Ho = 0.35). In *C. araguaiaensis* 42 SNPs were identified but the vast majority were at low frequency. These SNPs were evenly distributed across the gene with the exception of the LLR C terminal region, which had a peak of higher frequency SNPs (both synonymous and non-synonymous) (Ho = 0.60) (Figure 3.9).

In TLR2, in *C. maculifer* a single synonymous SNP was detected; occurring in *c.*82% of the sample population (Ho = 0.63). In *C. araguaiaensis* 114 SNPs were found across the population, again relatively evenly distributed across TLR2, with the majority at low frequency. However, one non-synonymous alternative base was present in all individuals within the population (either in a heterozygous or homozygous state). In this instance the proportion of haplotypes carrying this alternative base might be a more useful indicator of the SNP frequency. This analysis found that even accounting for four haplotypes per individual this alternative base was present in 75% of haplotypes (lower plot in figure 3.10). A further three synonymous SNPs were present at relatively high frequencies (>50% of individuals, Ho = 0.65, 0.89, 0.66 respectively) (Figure 3.10).

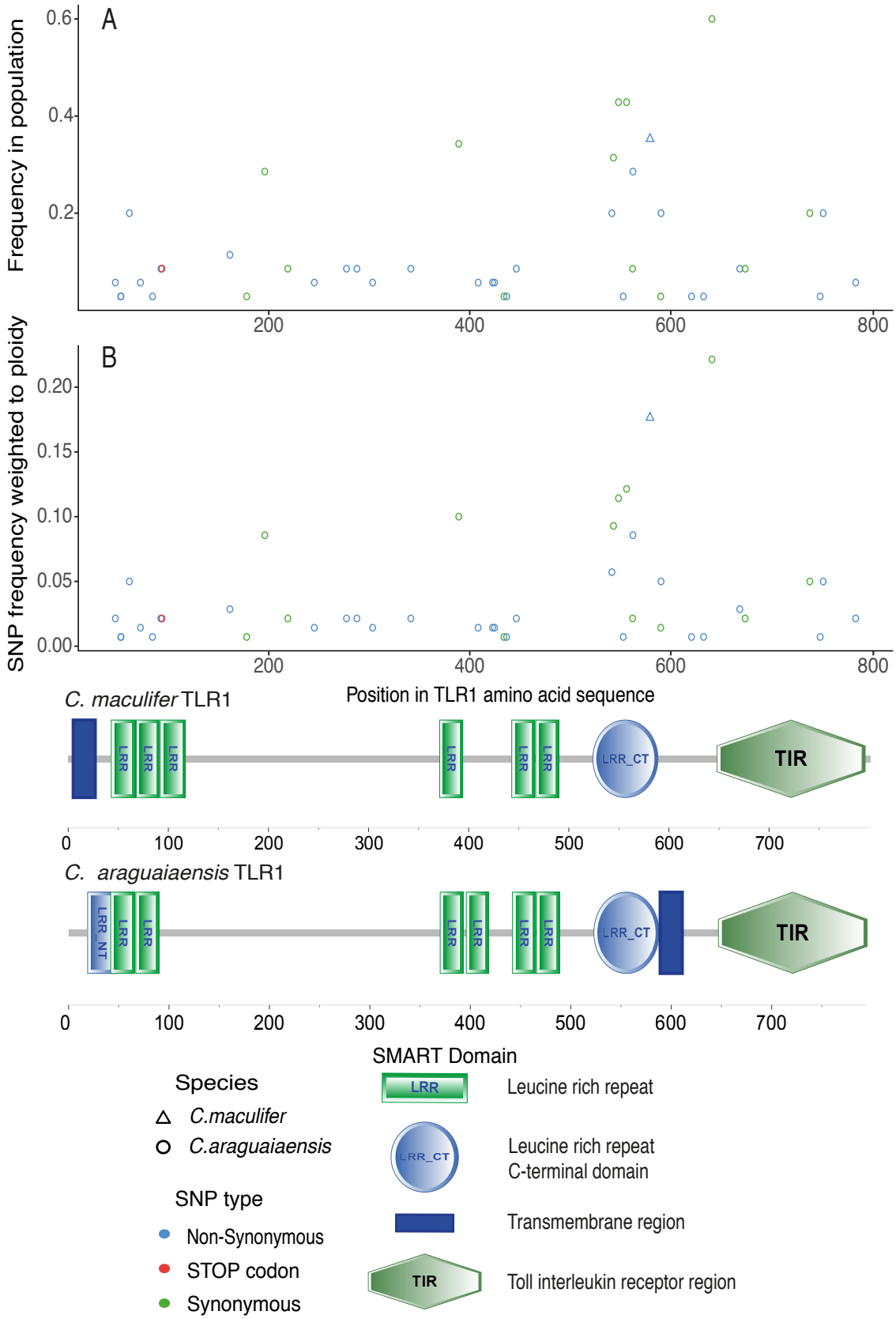


Figure 3.9: A: Frequencies of alternative bases in TLR1 across populations of C. maculifer and C. araguaiaensis. Counted across individuals regardless of ploidy status. B: Alternative base frequencies in TLR1 across populations of C. maculifer and C. araguaiaensis weighted according to proportional presence based on read depth. Assuming diploidy in C. maculifer and tetraploidy in C. araguaiaensis. Both plots are aligned to the protein domain output derived by SMART analysis for TLR1 for each species.

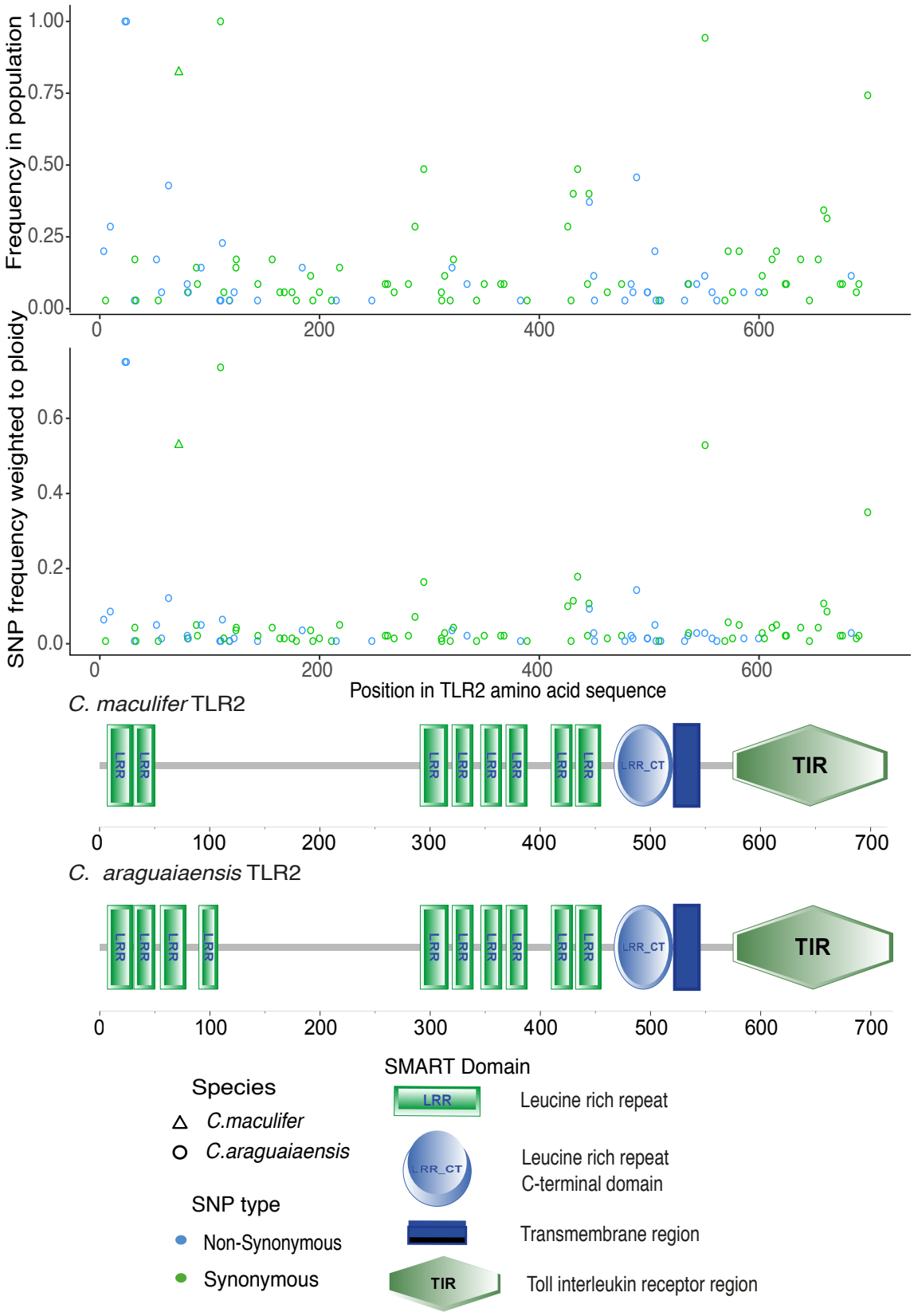


Figure 3.10: A: Frequencies of alternative bases in TLR2 across populations of C. maculifer and C. araguaiaensis. Counted across individuals regardless of ploidy status. B: Alternative base frequencies in TLR2 across populations of C. maculifer and C. araguaiaensis weighted according to proportional presence based on read depth. Assuming diploidy in C. maculifer and tetraploidy in C. araguaiaensis. Both plots are aligned to the protein domain output derived by SMART analysis for TLR2 for each species.

3.4 Discussion

In this chapter, genetic diversity at two TLR genes across two species of *Corydoras* catfish with different ploidy states was examined in an attempt to establish whether WGD was linked to increased TLR gene diversity, and the potential advantages this might confer. The results highlighted a number of patterns within and between the two species assessed.

C. araguaiaensis populations exhibited significantly diversity across both TLR genes then *C. maculifer*. Indeed, individual SNP profiles were also frequently unique within the *C. araguaiaensis population* but not in the *C. maculifer* population. Haplotype count analyses indicated a maximum of two haplotypes across both TLRs in individuals of *C. maculifer.* In contrast, SNP ratio analysis suggested the four haplotype exist in most *C. araguaiaensis* individuals, while QualitySNPng short-range phasing indicated that some individuals of *C. araguaiaensis* might have five TLR2 haplotypes. The predicted structures of both TLR1 and TLR2 were relatively similar between *C. maculifer* and *C. araguaiaensis*, the only major difference was in the placement of the transmembrane region in TLR1 in *C. maculifer*. Finally, other then a slight peak in the LRR C terminal region in TLR1, there were no clear patterns in the distribution and frequency of SNPs across functional regions of the TLR genes.

3.4.1 Higher diversity among individuals and across the population of C. araguaiaensis

More SNPs were found in the TLR genes of the polyploid, *C. araguaiaensis* than in the diploid *C. maculifer.* The genetic diversity of diploid *C. maculifer* was found to be exceptionally low, exhibiting comparable or lower numbers of SNPs per TLR to, for example, previously bottlenecked and/or threatened bird populations (Grueber *et al.*, 2015; Gilroy *et al.*, 2017). It may be that *C. maculifer* populations have undergone a population bottleneck in their recent evolutionary past. However, this does not detract from the relatively high numbers of SNPs observed in *C. araguaiaensis*. An observation which supports the theory that organisms of higher ploidy are more subject to genetic drift and mutation due to additional gene copies being freed from selection. In addition, there was a high degree of difference between SNP profiles within the *C. araguaiaensis* population, where individual often exhibited unique profiles. This may be indicative of ploidy related dilution of selection, whereby any advantageous haplotypes present are likely to be found at very low frequency within the population.

The high degree of potential functional diversity shown in individual SNP profiles in *C. araguaiaensis* may support the idea that individuals benefit from advantages linked to heterozygous advantage and negative frequency dependence in this species. With the

exception of four individuals, all *C. araguaiaensis* individuals were functionally unique at both TLR loci, allowing them the potential genetic space and variation to carry and benefit from an advantageous haplotype. In addition, the high level of variation between SNP profiles and the frequency of unique SNP profiles indicates a greater probability that individuals carry rare haplotypes. Rare haplotypes could prove advantageous in immune genes in relation to host parasite interactions due to mechanisms surrounding negative frequency dependence (the idea that pathogens will be under selection to develop evasion mechanisms for commonly occurring immune haplotypes).

The polyploid origin of *C. araguaiaensis* is not yet understood. It could have arisen as part of a hybridisation event (allopolyploid) or through mechanisms of chromosome retention during gametogenesis or fertilisation (autopolyploid) (Mable, Alexandrou and Taylor, 2011). The predicted effects of these two different mechanisms of WGD on gene diversity are subtly different. Allopolyploids might benefit from fixed heterozygosity through the enforced pairing of homologous progenitor chromosomes, preventing inter-genomic recombination and effectively maintaining heterozygosity across the generations (Comai, 2005). Evidence from allopolyploid cotton indicates that duplicated genes evolve independently from one another (Cronn, Small and Wendel, 1999). In addition, potentially deleterious recessive alleles are masked in both ploidy scenarios. An *Aa* heterozygote diploid would be expected to produce 1/4 *aa* offspring, an autopolyploid *AAaa* would produce between 1/36 and 1/22 *aaaa* homozygotes, and a allopolyploid *AaAa* would produce 1/16 *aaaa* homozygotes (Comai, 2005). There are no data on the chromosomal behaviour or genomic origin of *C. araguaiaensis*. Therefore, it is not possible to say if any of these allo- or autopolyploid signatures are coming into play across the species examined here, as both would be predicted to increase genetic diversity and heterozygosity but in subtly different ways.

3.4.2 Haplotype retention

Both QualitySNPng short phasing and SNP ratio histograms showed that *C. maculifer* had a maximum of two haplotypes per individual. In contrast, in *C. araguaiaensis* manually validated short-range phasing revealed that some individuals (n = 4) had up to five haplotypes in TLR2, although the remaining individuals (n = 31) had up to four haplotypes at both loci. When both loci were combined, SNP read ratio histograms indicated that *C. araguaiaensis* had four haplotypes across the population, which is unsurprising given the relatively low number of individuals carrying five haplotypes and these were at a single locus.

This haplotype quantification analyses provided evidence that *C. maculifer* is diploid and *C. araguaiaensis* is tetraploid. It also confirmed that four copies of TLRs 1 and 2 are likely to

have been maintained in *C. araguaiaensis*, despite the trend for rapid re-diploidisation via gene fractionation following WGD. However, the individuals that exhibited five haplotypes remain anomalous. Possible explanations for the fifth haplotype include; remnants from a more ancient WGD event prior to the one that induced tetraploidisation, tandem duplication in the TLR, and PCR or sequencing error. Ancient WGD events are thought to have occurred several times during the evolutionary history of the *Corydoradinae*, i.e. during the diversification of lineage four and possibly again at the base of lineage nine (Marburger *et al.*, 2018). It is conceivable that the fifth haplotype present in some *C. araguaiaensis* individuals may be a remnant from one of these events that has not yet been lost. Tandem duplications are single gene duplications that occur either by unequal crossing over during meiosis or by retrotransposition (Temperley *et al.*, 2008). Tandem duplications have been documented in the TLR gene family in other species, though they don't appear to be as common as in other gene families (i.e. MHC) (Temperley *et al.*, 2008). Consequently, tandem duplication may be another mechanism that could explain the presence of the fifth TLR2 haplotype in *C. araguaiaensis*. The process of PCR is universally used for amplicon sequencing; however, it can result in errors. This can't be completely ruled out as a possible explanation for the fifth haplotype observed in *C. araguaiaensis* without being able to phase haplotypes and compare similarities between individuals. However, no such errors were identified in *C. maculifer*, which may have been expected if errors were responsible for the additional haplotypes in *C. araguaiaensis*. Finally, the fifth haplotype could be the result of an erroneous haplotype call in QualitySNPng, however as all haplotypes were validated manually this is extremely unlikely.

3.4.3 Structural variation and the distribution of SNPs across TLRs

Structural protein analysis identified a transmembrane region at the N terminus of TLR1 in *C. maculifer,* whereas the transmembrane region in TLR1 of *C. araguaiaensis* was located in the 'normal' position between the LLR-CT region and the TIR domain. TLR1LB and some variants of TLR7 in chickens, *Gallus gallus domesticus*, also have a transmembrane region at the N terminus (Temperley *et al.*, 2008). However this positioning of a transmembrane region is not common and it is possible that the SMART algorithm falsely interpreted a hydrophobic region as a transmembrane region (Temperley *et al.*, 2008). SMART analysis looks for signatures of different protein domains by comparing amino acid sequences back to protein databases (Letunic and Bork, 2018), some of these signatures might be expected to be more specific than others. Transmembrane regions are associated with hydrophobic signatures, which is a relatively general association, and may therefore lead to misidentifications (Temperley et al., 2008).

SNP distribution and frequency plots showed that most SNPs were at low frequency across *C. araguaiaensis* and relatively evenly distributed across both TLR genes. A peak in SNP frequencies between 500aa and 600aa, which corresponded to the LLR-C terminal region in TLR1 (part of the PRR domain), was detected in *C. araguaiaensis*, and also corresponded to the single SNP identified in TLR1 in *C. maculifer*. One SNP coding for a stop codon was detected in *C. araguaiaensis* in TLR1 in three individuals suggesting some copies of the locus may be non-functional.

The even distribution of low frequency SNPs across both TLR genes may support the freedom from selection hypothesis, under which additional gene copies are free to drift in frequency and accrue mutations. Moreover, selection on beneficial mutations may be weak and deleterious mutations masked due to haplotype number thus explaining why most of these SNPs are at low frequency within the population. The presence of a stop codon among the non-synonymous SNPs identified could indicate silencing of additional haplotype copies, however this stop codon was only detected in 3 individuals (8.6% of the sample). The peak of higher frequency SNPs over the LLR-C terminal domain in both species of catfish may indicate that diversity in this region is beneficial to the PRR domain.

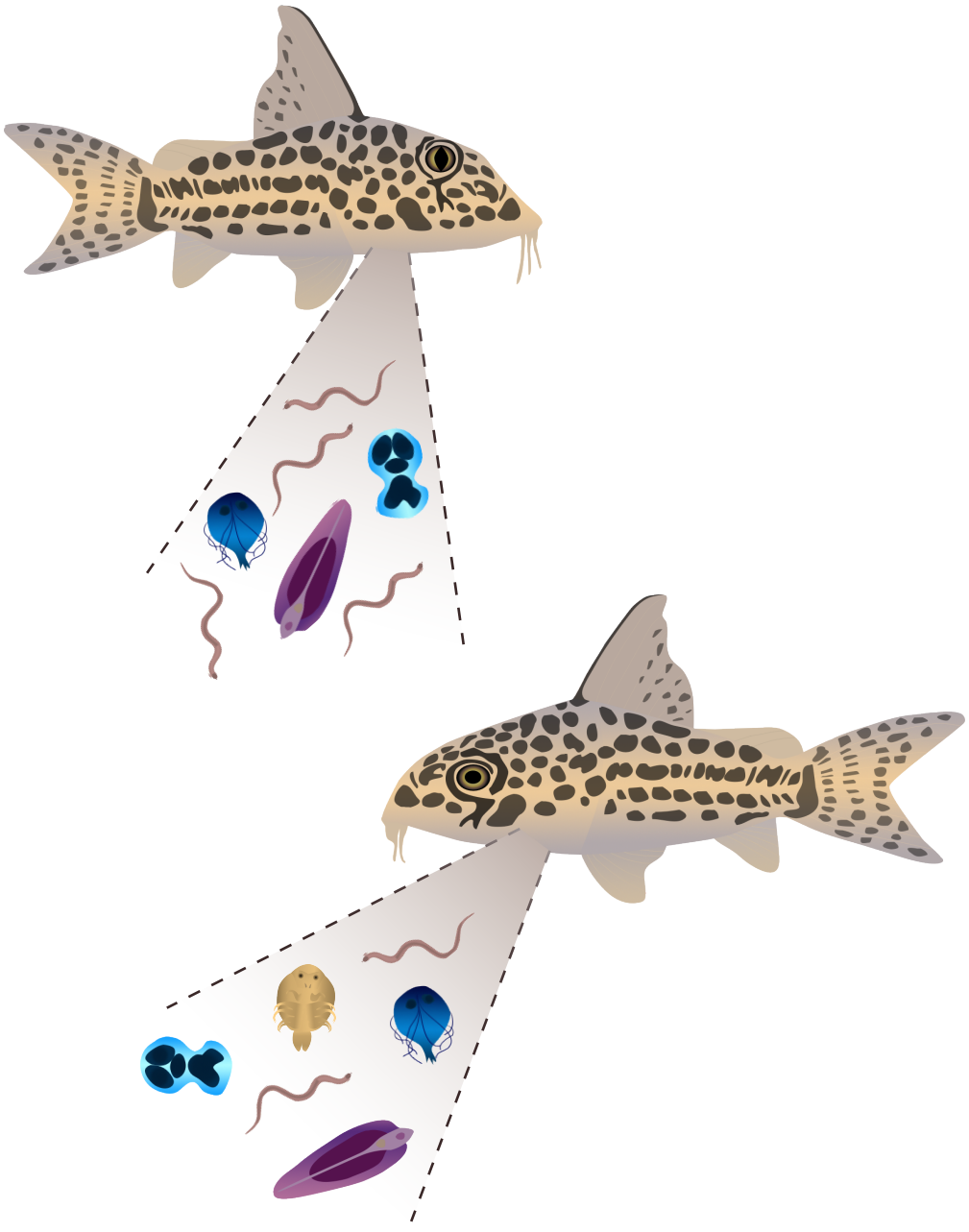
The stop codon within TLR1 in some individuals suggests that at least some of the additional TLR copies, in some individuals, are not expressed. However, this study measured genetic diversity but not expression, so we have no other evidence of differential expression. Expression patterns in polyploid organisms are complex, duplicate genes may be up or down regulated, truncated or silenced altogether (Adams and Wende, 2005; Fink *et al.*, 2016). Gene regulation in polyploids remains poorly understood but, in addition to genetic modifications, mechanisms associated with epigenetic pathways and RNA-mediated pathways have been implicated (Chen, 2007). When the full transcriptome of leaf cells from the recent allopolyploid *Glycine dolichocarpa* were compared to progenitor species, the transcriptome was found to be c.1.4 fold larger than either progenitor, and 70% larger than the sum of both progenitors combined. However, the allopolyploid genome was smaller than the size of both progenitor genomes combined (only 94.3% the size of the sum of both progenitor genomes). So although the allopolyploid transcriptome was greater than either progenitor its genome had reduced at a greater rate (Coate and Doyle, 2010). It is therefore possible that although the TLR genes may be more diverse, they may be being either up or down regulated, in some cases silenced. The current data cannot shed further light on this.

3.4.4 Conclusion

Overall *C. araguaiaensis* exhibited greater diversity then *C. maculifer*, and appears to have retained at least four copies of the TLR genes. However, the current data could not be phased, so full length haplotypes could not be identified. This hampers further analysis on selection and prevents examination of the frequency of haplotypes within each population. Furthermore, although haplotype number could be estimated these data do not provide any indication as to whether all copies are transcribed and functional. One SNP was identified as coding for a stop codon but was only present in three individuals and was the only indication of haplotype silencing in these data.

Chapter 4:

Parasite community comparisons between diploid and polyploid *Corydoras* catfish hosts



4.1 Introduction

Parasites are commonly defined as organisms that live in or on another organism (i.e. the host) feeding on it, causing it a degree of harm and showing some level of adaption to it (Poulin, 2007). Host-parasite interactions can be highly complex and may be affected by a number of biotic and abiotic factors including, host age, size, behaviour, physiology, diet, immunology, and habitat (Ryce, Zale and MacConnell, 2004; Khan, 2012; Lester and McVinish, 2016). These relationships are also complicated by the complex life histories and specificities of the parasites themselves; many parasites undergo multiple life stages, which can be associated with different host species, terminating in a definitive host in which the parasite reaches sexual maturity (Poulin, 2007). These combined factors make exploring host parasite interactions challenging.

One key factor that will influence host/parasite interactions is host immunity (Alvarez-Pellitero, 2008). The immune systems of animals serve as a vital defence against invading pathogens (including parasites). Immune systems are broken down into innate and adaptive mechanisms and further subdivided into a range of specialised cells and proteins adapted for specific roles within host defence (Takeda and Akira, 2005). A vital function of this system is the recognition of diverse foreign antigens (Takeda and Akira, 2005). Pathogen recognition receptors (PRRs) are part one of the primary defence mechanisms for recognising pathogen associated molecular patterns (PAMPs) (Alvarez-Pellitero, 2008). In fishes one of the major classes of PRR are the Toll-like receptors (TLRs) (Aoki and Hirono, 2006).

The TLRs are a class of type 1 transmembrane proteins that are encoded to recognise extracellular PAMPs and initiate an immune response (Zhao *et al.*, 2013). Once activated TLRs are thought to play direct roles in the inflammatory response and indirectly influence adaptive responses through the regulation of antigen presentation on dendritic cells (Salaun, Romero and Lebecque, 2007). Diversity in PRRs is thought to be advantageous and the genes that encode TLRs are highly polymorphic (Netea, Wijmenga and O'Neill, 2012). Several studies have found evidence for potential associations between specific polymorphisms in TLRs and disease susceptibilities (Skevaki *et al.*, 2015).

Variation in and among immune genes, and therefore in the proteins they encode, is essential to detect such a range of antigens and high levels of variation are well documented in immune gene families such as the Major Histocompatibility Complex (MHC) (Zinkernagal and Doherty, 1974; Hill, 1999; Phillips *et al.*, 2018). Whole genome duplication (WGD) events represent potential mechanisms for rapidly increasing genetic diversity in host species (Mable, Alexandrou and Taylor, 2011). However, empirical evidence supporting the mechanics of this process is lacking, and knowledge in this area is largely reliant on theory. Polyploid individuals

are more likely to carry variation at a given locus and may therefore also have a higher probability of carrying rare haplotypes (Otto and Whitton, 2000; King, Seppälä and Neiman, 2012). This makes them more likely to benefit from mechanisms such as heterozygote advantage and negative frequency dependence (Doherty and Zinkernagel, 1975; Slade and McCallum, 1992; King, Seppälä and Neiman, 2012). Polyploid individuals may also benefit from dosage dependent effects due to the increased potential number of copies of a haplotype that may be translated allowing them to produce greater quantities of immune related proteins (King, Seppälä and Neiman, 2012).

The Corydoradinae are a species rich subfamily of Neotropical catfishes found across large parts of South America (Fuller and Evers, 2005). A recent phylogenetic analysis identified nine lineages within this subfamily and single or multiple WGD events in the evolutionary history of some lineages (Alexandrou *et al.*, 2011; Marburger *et al.*, 2018). Uniquely, species belonging to the Corydoradinae often form mixed sympatric communities of up to three species, often from different genetic lineages and with differing genome sizes (Alexandrou *et al.*, 2011). These mixed communities have a propensity towards Müllerian mimicry and evidence has also been found for resource partitioning (Alexandrou *et al.*, 2011). The majority of Corydoradinae species are benthic scavengers and occupy trophic levels associated with omnivorous detritivores (Nijssen, 1970). Coexisting Corydoradinae species may have differing snout morphology, and stable isotope analysis of dietary content suggested a dietary segregation between species. Larger, longer snouted species were found to occupy lower trophic levels (lower $\delta^{15}N$) then smaller, shorter snouted species (Alexandrou *et al.*, 2011).

The Corydoradinae community in the Araguaia River in the state of Mato Grosso, Brazil, consists of three species, *Corydoras maculifer* (lineage 1, diploid), *Corydoras araguaiaensis* (lineage 9, putative tetraploid) and an un-described *Corydoras sp*. (lineage 8) (Alexandrou *et al.*, 2011; Marburger *et al.*, 2018). The three species form a mimicry ring (Alexandrou *et al.*, 2011), however little is known about the very rare and currently un-described lineage 8 species and this species is not included in this chapter. Within this community *C. maculifer* is the larger species with a longer snout and occupying a lower trophic level, while *C. araguaiaensis* is smaller with a shorter snout and occupies a higher trophic level (Alexandrou *et al.*, 2011). Because these two species coexist in the same area, and have similar dietary habits it is likely that they will have been exposed to similar parasitic communities.

4.1.1 Aims and objectives

Here we investigate the potential relationships between immune gene diversity, parasite burden and community abundance of two sympatric *Corydoras* catfish species, *C. maculifer*

(lineage 1, diploid) and *C. araguaiaensis* (lineage 9, putative tetraploid). We aim to explore differences in their parasite burdens and examine potential links with differences in their TLR immune gene composition. Chapter 3 found significantly higher diversity in TLR1 and TLR2 in the putative tetraploid *C. araguaiaensis*. Diversity in immune genes is thought to be advantageous so we hypothesised that parasite burden would be reduced in the tetraploid population when compared to the diploid population of *C. maculifer*.

4.2 Methods

4.2.1 Host sampling

Individuals of *C. maculifer* (n=20) and *C. araguaiaensis* (n=41) were collected from the wild from the same location in the Rio das Mortes drainage of the Araguaia River in Mato Grosso state, Brazil by MIT, CO (2012 and 2015) and EB (2015), euthanized by anaesthetic overdose and stored individually in 100% ethanol. The 2015 cohort were dissected on site to remove liver, gonads, stomach and digestive tracts, these organs were preserved separately to the rest of the sample in 100% ethanol. This measure was undertaken to maximise DNA preservation of internal parasites.

4.2.2 Parasite extraction, identification and enumeration

Individual samples of *C. maculifer* and *C. araguaiaensis* were screened for parasites. Each individual was screened externally prior to body cavity, stomach, digestive tract and liver tissues being dissected separately and examined for parasites. Parasites found were grouped by host sample, the tissue they were found in, and by morphological similarity. Parasite samples were counted and stored in 100% ethanol. Identification at this stage was largely descriptive and very broad (i.e. nematode, encysted, isopod etc). Parasite samples were photographed where possible prior to DNA extraction (see Appendix).

4.2.3 Parasite DNA extraction, PCR amplification and sequencing

The majority of parasites extracted were nematodes. DNA was extracted from individual nematode samples using a modified version of the salt extraction protocol developed by Sunnucks & Hales 1996 and Aljanabi & Martinez 1997. Individual nematodes were incubated at 55^{o}C overnight in 50µl of digestion buffer (30mM Tris-HCL pH8, 10mM EDTA, 1% SDS) and 2µl of proteinase K. Following digestion 20µl of 5M NaCl was added to each sample before being centrifuged at 13,000 rpm for 5 minutes. Supernatant was then transferred to a new tube and twice the transferred volume of 100% ice-cold ethanol added prior to incubation at -80^{o}C for 20 minutes. Samples were then centrifuged for 30 minutes at 13,000rpm and supernatant discarded. Samples were washed with 70% ethanol, re-centrifuged at 13,000rpm for 5 minutes and dried at 35^{o}C before re-suspension in 20µl of dH_2O.

PCR amplifications used the method outlined by Prosser et al. (2013) which involves a universal primer 'cocktail' for amplification of CO1 in all nematode species. This cocktail combined three forward (Nem_PCR_1_Fw, Nem_PCR_2_Fw, Nem_PCR_3_Fw) and three

reverse (Nem_PCR_1_Rv, Nem_PCR_2_Rv, Nem_PCR_3_Rv) primers mixed in equal concentrations. For PCR amplification, 0.25µl of 10µM forward and reverse primer 'cocktail' (Table 4.1), 12.5µ of 2x PCRBIO Taq Mix Red (PCR Bio-systems) and 2µl of DNA were combined and made up to a final volume of 22µl with H_2O. PCR conditions were: initial denaturation of 94°C for 60s and then a secondary denaturation of 94°C for 40s, annealing temperature of 45°C for 40s, extension temperature of 72°C for 60s for 5 cycles. Followed by a tertiary denaturation step of 94°C for 40s, secondary annealing temperature of 51°C for 40s and secondary extension step of 72°C for 60s for 35 cycles proceeded by a final extension step of 72°C for 5 minutes. PCR products were cleaned using an ExoSap PCR clean-up protocol. These reactions combined 10µl of PCR product, 0.1µl of EXO1, 0.2µl of FastAp and 4.7µl of H_2O. Clean-up conditions were 37°C for 15 minutes followed by 80°C for 15 minutes. PCR products were visualised on ethidium bromide stained 0.8% agarose gels. Samples that produced clear clean bands of approximately the right size were submitted for Sanger sequencing with a separate forward sequencing primer (Nem_Seq, Table 4.1).

Table 4.1: Universal primers used for PCR amplification of CO1 in nematode parasites (using IUPAC nucleotide ambiguity codes)

Primer name	Forward	Reverse
Nem_PCR_1	TGT AAA ACG ACG GCC AGT CRA CWG TWA ATC AYA ARA ATA TTG G	CAG GAA ACA GCT ATG ACT AAA CTT CWG GRT GAC CAA AAA ATC A
Nem_PCR_2	TGT AAA ACG ACG GCC AGT GCC AGT ARA GAT CTA ATC ATA AAG ATA TYG GG	CAG GAA ACA GCT ATG ACT AWA CYT CWG GRT GMC CAA AAA AYC A
Nem_PCR_3	TGT AAA ACG ACG GCC AGT ARA GTT CTA ATC ATA ARG ATA TTG G	CAG GAA ACA GCT ATG ACT AAA CCT CWG GAT GAC CAA AAA ATC A
Nem_Seq	TGT AAA ACG ACG GCC AGT	

4.2.4 Data processing and analysis

Parasite prevalence is the proportion of individuals of each host species*, C. maculifer* and *C. araguaiaensis*, exhibiting a parasitic infection. Parasite intensity is the number of parasites found in an infected individual (hosts free of infection are removed from the analysis). Parasite abundance is a combination of parasite prevalence and intensity i.e. the number of parasites found in an infected individual including individuals free from infection. Prevalence intensity and abundance, broken down by host tissue type and host species, were represented by both medians and means and plotted.

Standard length of the host fish species was recorded as the distance between the tip of the snout and the base of the tail. This was plotted against total parasite abundance per host individual and regression lines fitted per host species based on a linear model. An analysis of co-variance (ANCOVA) was performed to test for an association between either standard length or host species, on parasite abundance. In order to test for differences in parasite abundance between the two-host species, a Poisson generalised linear model (GLM) was fitted using the GLM function in R (version 3.4.1). The number of parasites was modeled, as a function of species and host length was included in the model as an offset. As our count data were over dispersed, the model was refitted using a quasi-Poisson distribution, which allows the dispersion parameter to be estimated from the data. The parameter estimates from these GLMs were then used to predict parasite abundance per millimetre for host species and plotted. Multi dimensional scaling (MDS) was also performed on the parasite abundance data using the R package Vegan (version 2.5-2) and a Bray-Curtis distances matrix.

Parasitic nematode molecular data of the C01 gene was blasted against the National Centre for Biotechnology Information (NCBI) Genbank database and top hits recorded. This sequence data was then combined with pre-existing molecular data collected from parasitic nematodes of other species of host *Corydoras* and aligned using the MUSCLE alignment algorithm from within Geneious (version 9.1.8). Maximum likelihood phylogenetic trees were constructed from these alignments using IQ-TREE (version 1.5.5, Nguyen et al. 2015; Hoang et al. 2018) and were based on the best model fit identified by jModelTest using the Bayesian information criterion (BIC). These trees were constructed using 1000 ultrafast bootstrap replicates and visualised in FigTree (version 1.4.3).

To assess potential links between immune gene diversity and parasite load, parasite abundance was plotted against non-synonymous single nucleotide polymorphisms (SNPs) found in the immune genes TLR1 and TLR2. Non-synonymous SNPs were derived following amplicon sequencing of TLR1 and TLR2 in *C. maculifer* (n=17) and *C. araguaiaensis* (n=36). Raw reads were mapped back to consensus sequence assemblies of each TLR and SNPs were called

based on these alignments (see Chapter 2 and 3 for details). Regression lines were fitted per host species using a linear model and an ANCOVA was performed to look for potential associations between TLR diversity host species and parasite abundance. *C. maculifer* had very limited intra-population diversity with only one non-synonymous SNP across TLR1 and TLR2. Consequently, subsequent immune gene analyses were only conducted within *C. araguaiaensis*. Maximum likelihood phylogenetic trees for consensus sequences of TLR1 and TLR2 in *C. araguaiaensis* were constructed using IQ-TREE and were based on the best model fit under the BIC as identified by jModelTest. Trees were constructed using 1000 ultrafast bootstrap replicates prior to being visualised using Figtree. Trees were then made ultrametric prior to being coloured according to ranked parasite abundance using Phytools (version 0.6-44, Revell 2012) in R studio. A SNP association analysis was performed using non-synonymous TLR SNP data to determine genotype and parasite abundance per *C. araguaiaensis* individual to describe host phenotype. The package GenABEL (version 1.8-0, Aulchenko et al. 2007) was used to perform the association analysis assuming a Gaussian distribution and to construct Manhattan plots. All plots, with the exception of Manhattan plots, which were made by GenABEL, were produced using ggplot2 (version 2.2.1) within R studio (R version 3.4.1).

4.3 Results

A total of 20 *C. maculifer* (all the samples available) and 41 *C. araguaiaensis* individuals were screened for parasites. Parasites were broken down into broad taxonomic and host tissue categories and enumerated. Gills were assessed in a number of individuals of both *C. maculifier* (n=7) and *C. araguaiaensis* (n=9), but with almost no parasites found assessment of this tissue was curtailed and any data collected were not analysed. A number of *C. maculifer* (n=5) also exhibited a possible fungal infection, but because fungal colonies were so numerous and easily disturbed there was no way of confidently counting them. These colonies were removed from subsequent analysis.

4.3.1 Parasite prevalence, intensity and abundance

Parasite prevalence (i.e. the percentage of infected individuals) was similar between the two host species (Figure 4.1). The exception to this was external parasite prevalence, which was significantly higher in *C. maculifer* (40.0% of individuals) than in *C. araguaiaensis* (4.9% of individuals) (Fishers Exact Test, n = 61, $p < 0.05$).

Parasite intensity (the numbers of parasites infecting a host - excluding non-infected individuals), was higher in all tissues in the diploid host *C. maculifer,* but this was not significant (Figure 4.2). The lack of statistical significance in these data may be due to the small sample sizes of infected individuals (sum total n = 16 *C. maculifer* and n = 27 *C. araguaiaensis*) as observational evidence from Figure 4.2 suggests a clear trend of higher intensities in *C. maculifer* over *C. araguaiaensis*. Mean parasite intensity was generally higher then the median in both species as well suggesting a right skewed distribution.

Parasite abundance is the number of parasites found in both infected and un-infected host species (Figure 4.3). When accounting for un-infected individuals, differences between the two species were less evident, although overall abundance was higher in *C. maculifer*. As with parasite prevalence, external parasite abundance was significantly higher in *C. maculifer* (Wilcoxon test: n = 20 *C. maculifer* and 41 C*. araguaiaensis*, $\overline{X}$ = 1.4 in *C. maculifer* and 0.05 in *C. araguaiaensis*, w = 559, $p < 0.01$).

When plotted against the standard length of host species, parasite abundance increased with standard length in both species (Figure 4.4). However a significant positive correlation was only detected in *C. araguaiaensis* (Spearman's rank correlation coefficient; S = 7312.1, rho = 0.343, $p < 0.05$). When assessing covariance between standard length and host species the only factor to have a significant effect was standard length (ANCOVA; F = 10.96, df = 1, $p < 0.01$).

There was a general skew in size distributions of hosts with greater proportions of large *C. maculifer* and small *C. araguaiaensis* in the sample. Given the apparent positive relationship between size and parasite abundance a GLM was used to take account of size when looking at different parasite abundances between the two species. This was used to model predictions per millimetre length of host for each host species (Figure 4.5). These predictions indicated parasite abundances were significantly different for external parasites and sum total parasite abundances between the two host species irrespective of size (GLM *C. araguaiaensis*; estimated Std = -3.22, Error = 0.95, t value = -3.40, $p < 0.01$ and estimated Std = -0.72, Error = 0.30, t value = -2.39 $p < 0.05$ respectively).

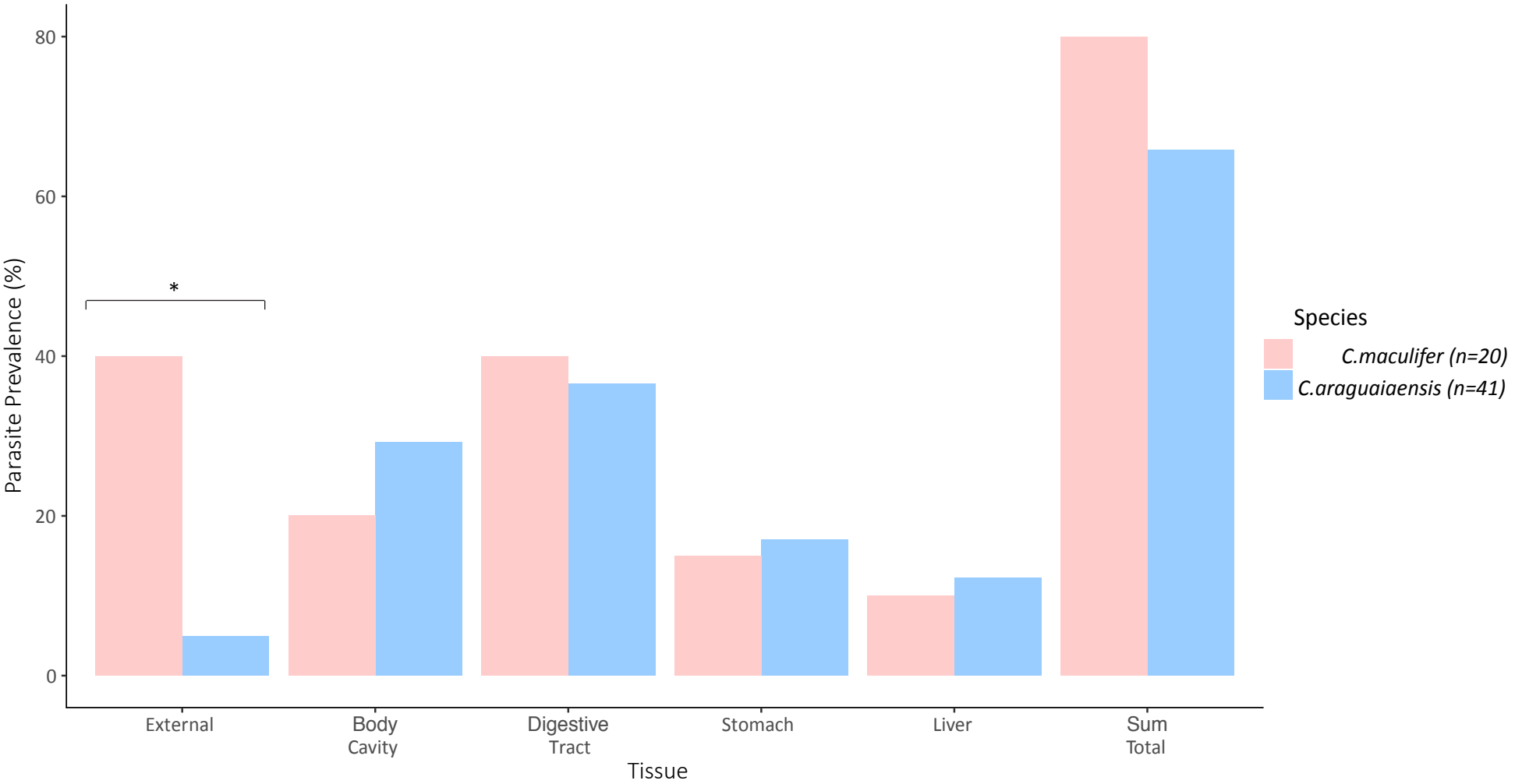


Figure 4.1: Prevalence (i.e. the proportion of infected hosts) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. Only external parasite prevalence was significantly different (Fishers exact test $p < 0.05$)

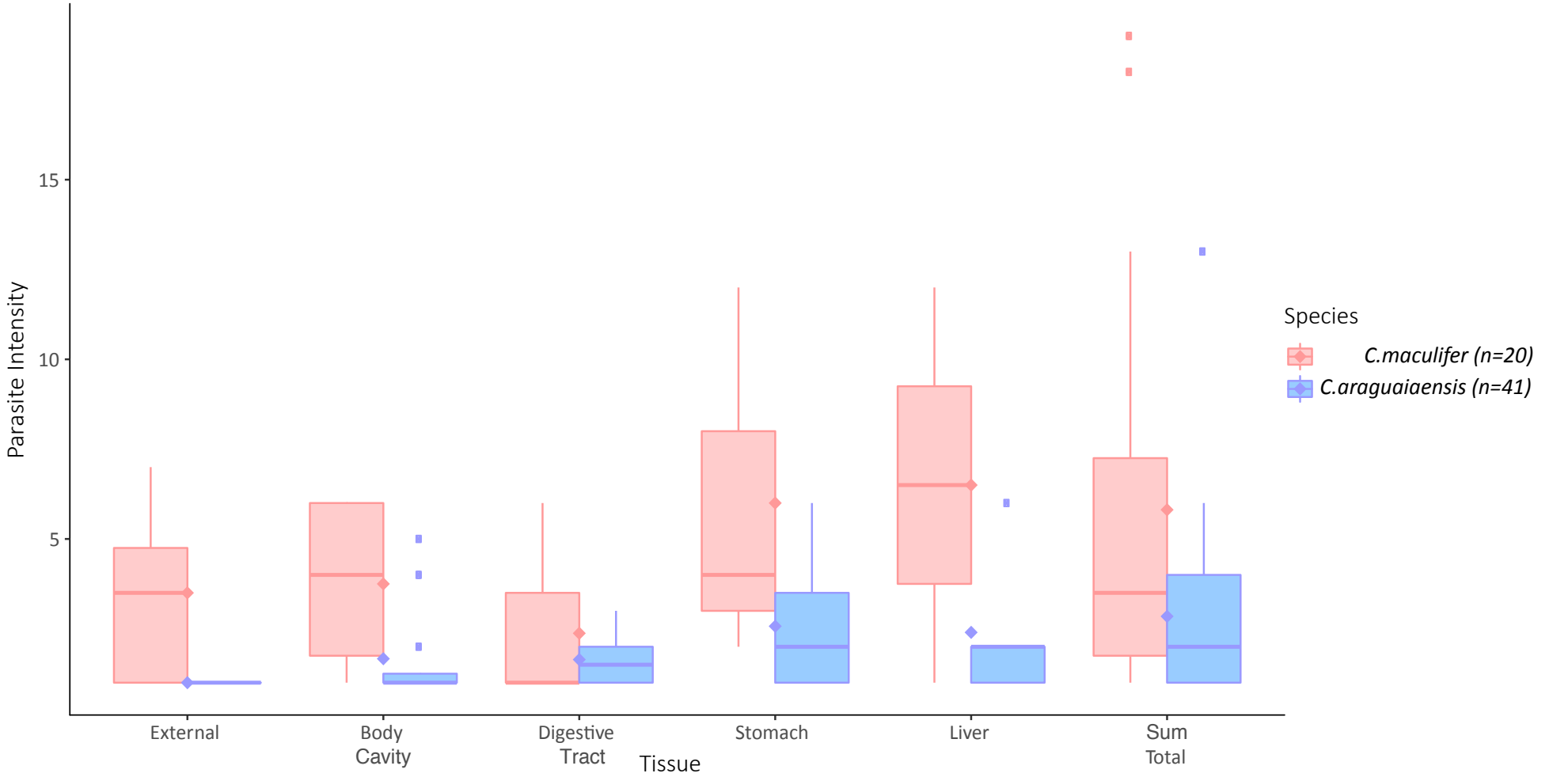


Figure 4.2: Intensity (i.e. the number of parasites per infected host) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. The central line in the box plots indicates the median and diamonds between boxes represent means. No significant differences were detected with either Moods Median Test, or Wilcoxon Test on averages or Fishers method of combining p values

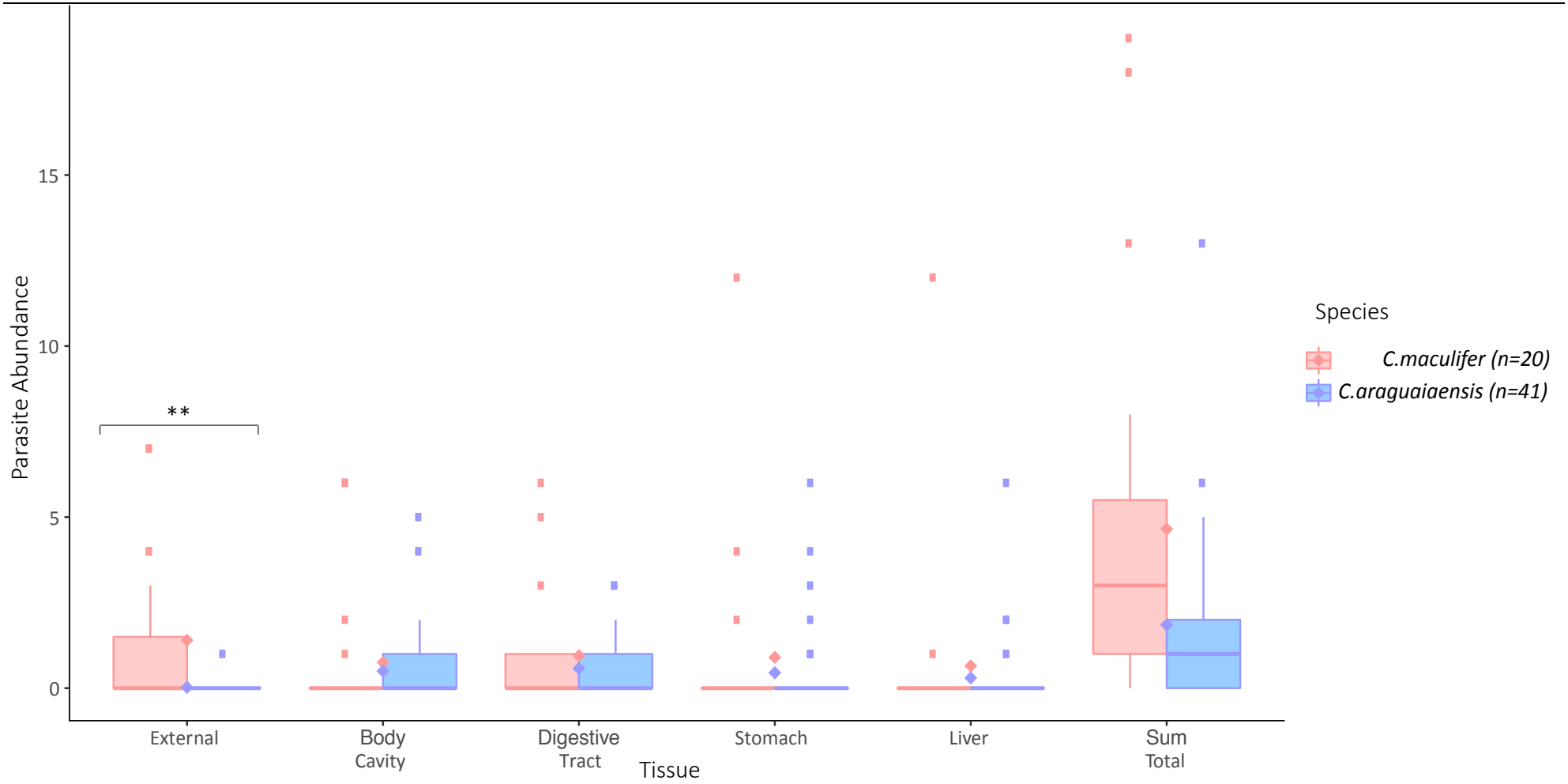


Figure 4.3: Abundances (i.e. the number of parasites per host) of parasites between two species of Corydoras catfishes, C. maculifer (diploid, n=20) and C. araguaiaensis (putative tetraploid, n=41), split according to tissue. The central line in the box plots indicates the median and diamonds between boxes represent means. Only mean external parasite abundance was significantly different (Wilcoxon test $p < 0.01$)

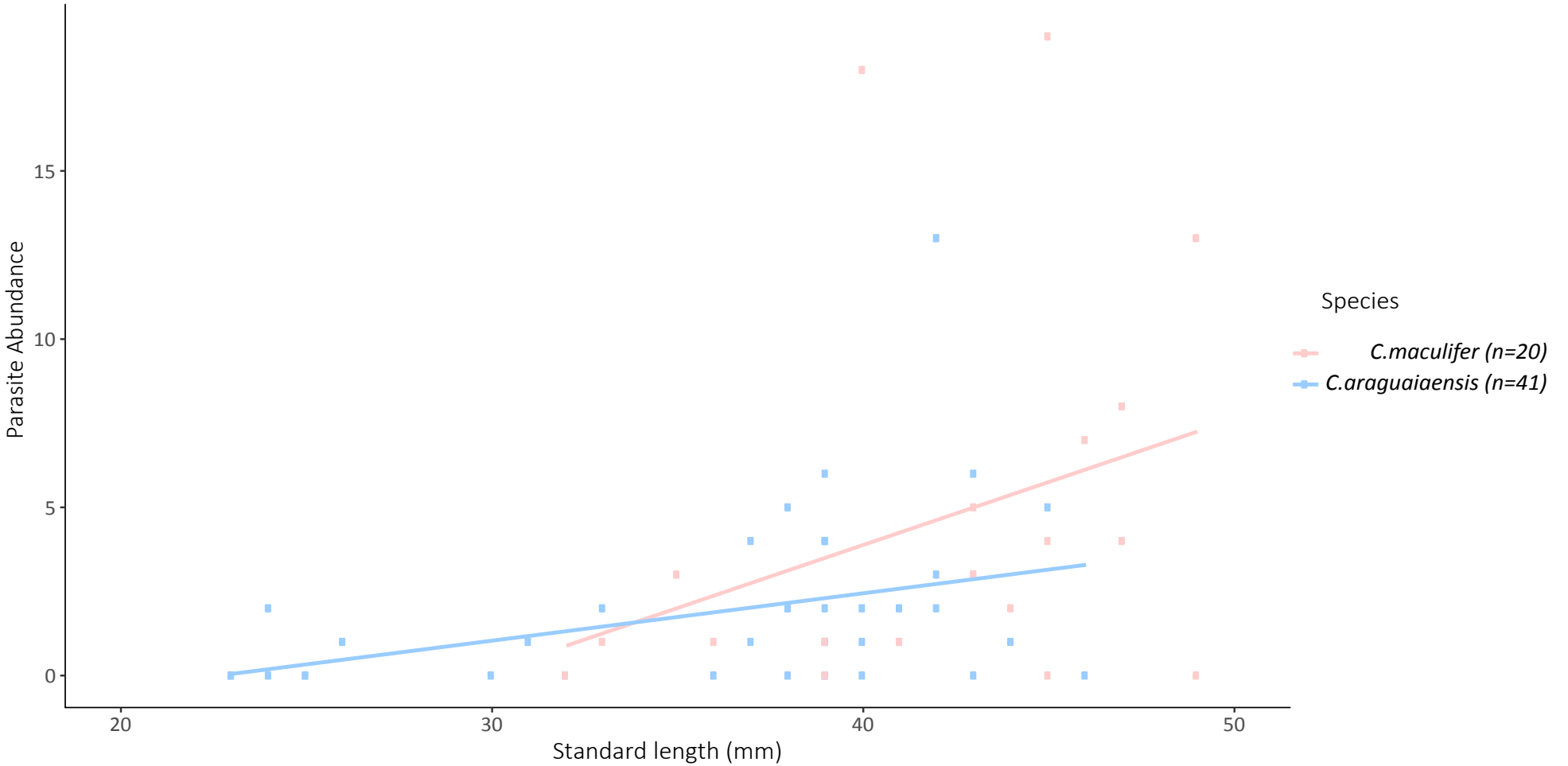


Figure 4.4: Standard length of host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), plotted against overall parasite count. ANCOVA showed standard length had a significant effect ($p < 0.01$) but there was no significant effect of host species on parasite count

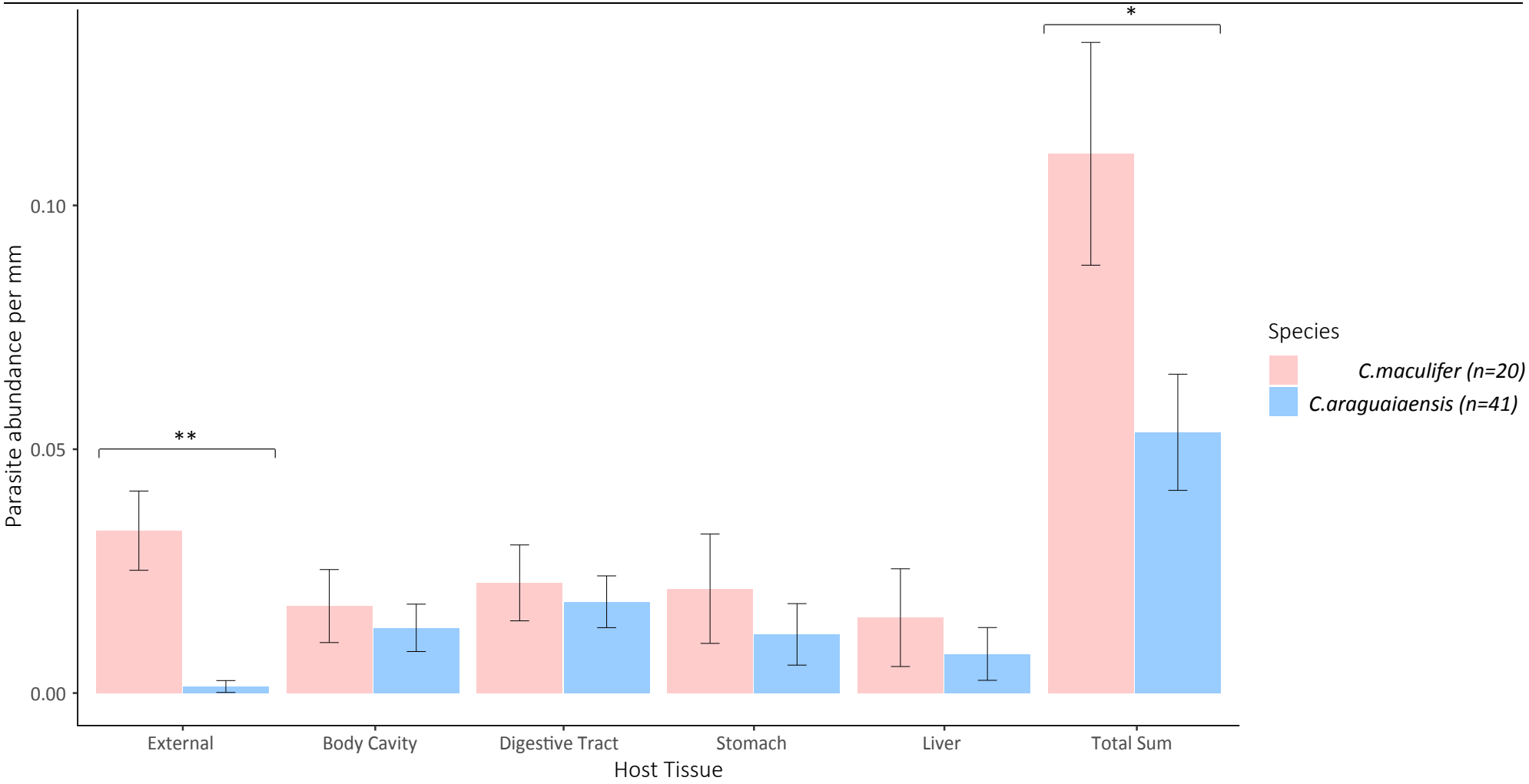


*Figure 4.5: Predicted parasite abundances (number of parasites per host) per millimetre of host plotted according to host species and infected tissue type. Predictions based on the outcomes of a general linear model (GLM) assuming a quasi-Poisson distribution to account for over dispersion (** = p < 0.01, * = p < 0.05 according to the GLM)*

4.3.2 Parasite community analysis

Parasites were sub-divided by morphology and host tissue type prior to intensity and abundance being calculated. When comparing parasite community intensities, cysts were found in greater intensities in *C. maculifer* in all tissues with the exception of liver tissues (Figure 4.6). Cysts were only found in the livers of *C. araguaiaensis*. A number of low intensity parasites, including *Isopoda* and *Acanthocephala* along with several unidentified probable parasites were found in *C. araguaiaensis* but were completely absent from *C. maculifer*. Parasitic nematodes were present in both host species but were generally found at higher intensities in *C. maculifer*.

Parasite community abundances, which included uninfected individuals (Figure 4.7), identified the majority of parasite occurrences as outliers because the majority of hosts did not share specific parasite/tissue infections. Differences in intensities and abundances, although apparent, were not significant. As with the overall parasitic intensities this might be because sample sizes of infected individuals were small. This is supported by Figure 4.7, which shows that the majority of infected host samples as outliers and demonstrates how small samples of infected hosts were within the population as a whole.

The collective parasite community data was used to produce an MDS plot based on a Bray-Curtis distance matrix. Host species largely overlapped without any clear segregation in parasitic communities (Figure 4.8).

A total of 29 nematode PCR amplifications were sequenced (2 from *C. maculifer* hosts and 27 from *C. araguaiaensis* hosts). One sample failed to provide a readable sequence and was eliminated from subsequent analyses. The resulting sequence blast results show that most nematodes were of the genus *Baylisascaris* (n=21), with the remaining nematodes belonging to the *Contracaecum* (n = 3), *Toxocara* (n = 2), *Camallanus* (n = 1) and *Ortleppascaris* (n = 1) genera. These parasite sequences were aligned with previously acquired parasite sequences from other *Corydoras* hosts (MIT unpublished) and used to construct a phylogenetic tree to explore parasite community structure (Figure 4.9). There was a regular overlap between the parasite branches and the *Corydoras* hosts they were associated to suggesting shared parasites between different host species.

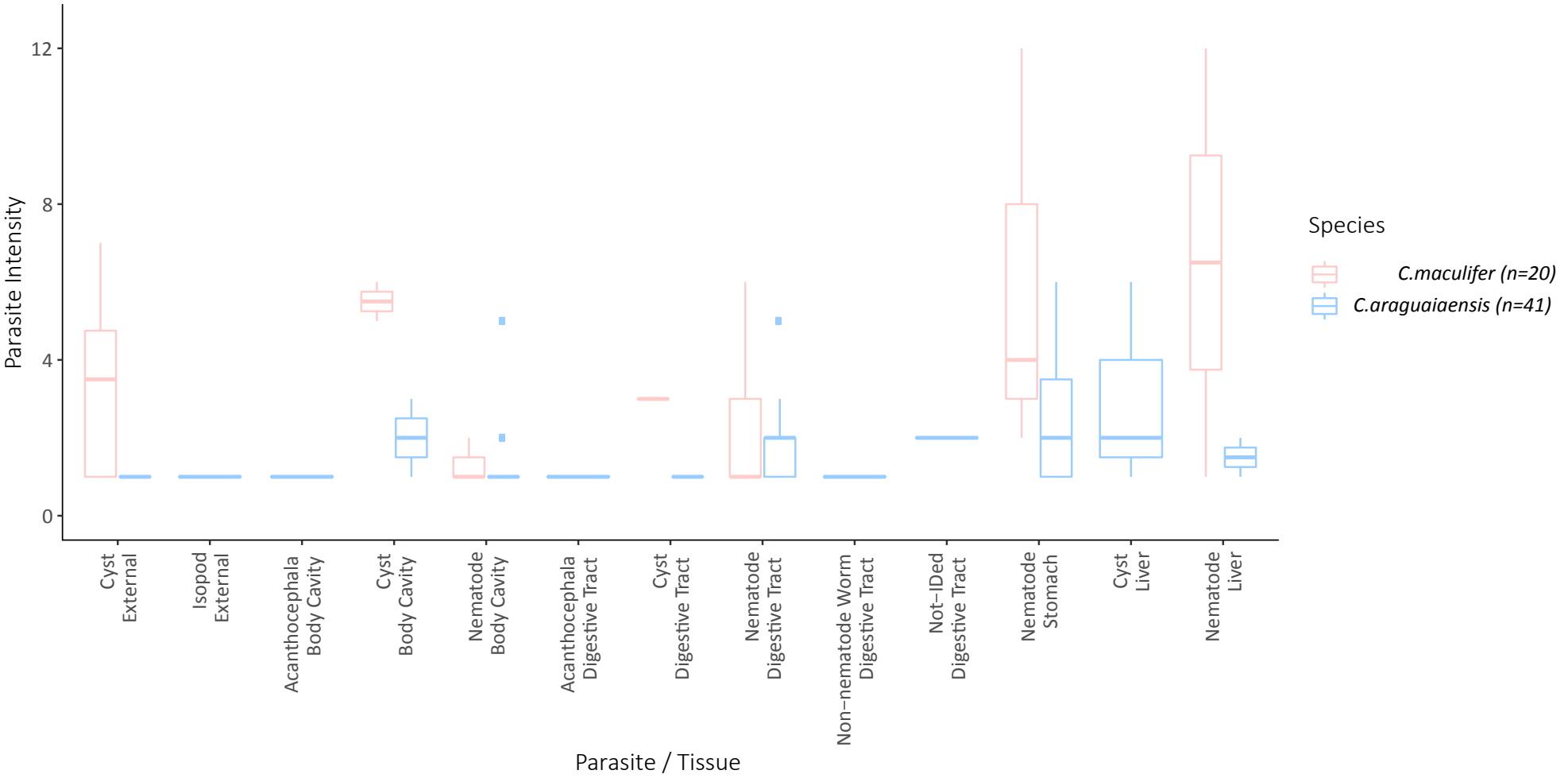


Figure 4.6: Intensity of parasites between two species of Corydoras catfishes, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), split according to tissue and parasite morphology. No significant differences were detected with Wilcoxon Test

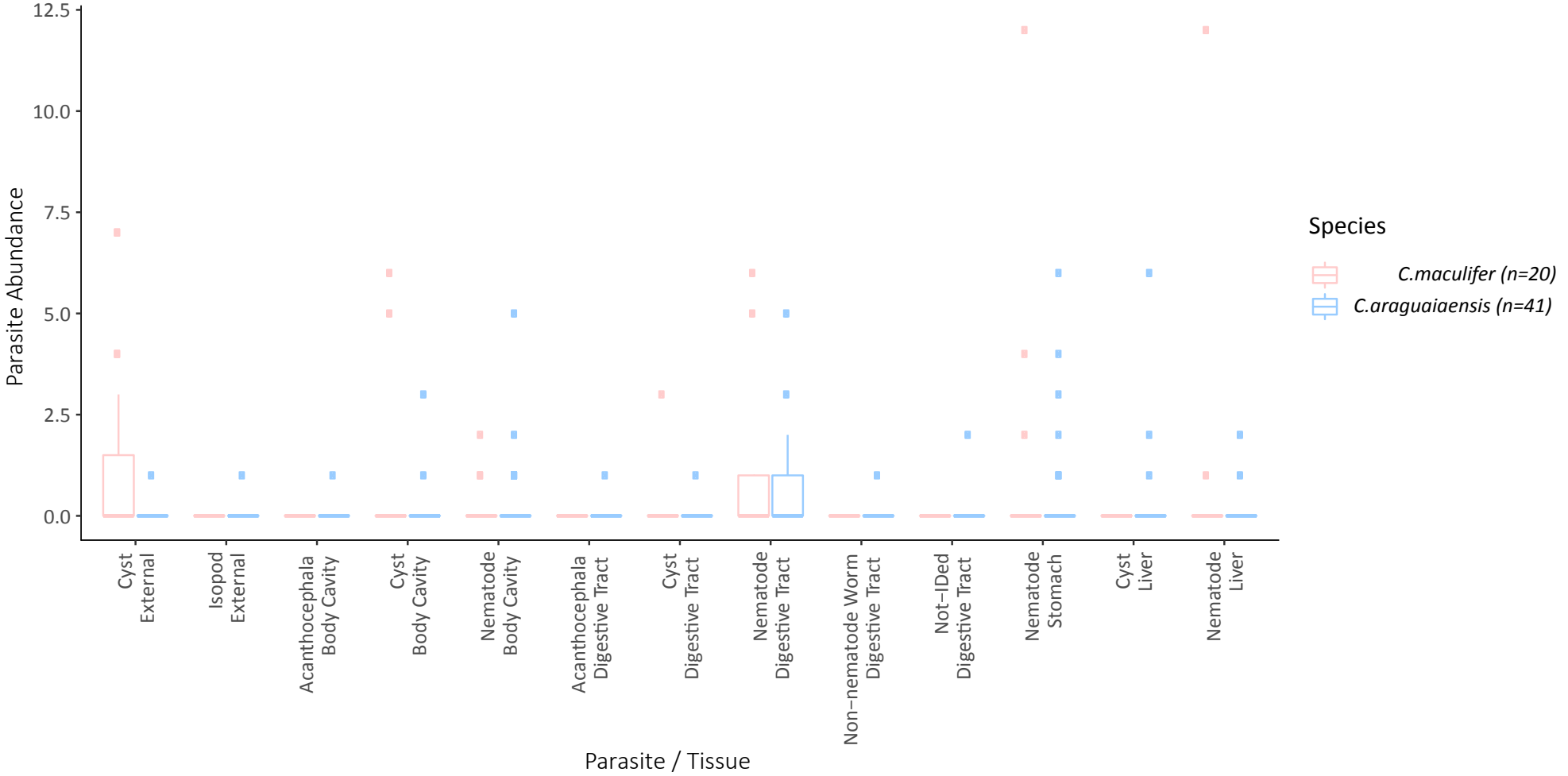


Figure 4.7: Abundances of parasites between two species of Corydoras catfishes, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), split according to tissue and parasite morphology. No significant differences were detected with Wilcoxon Test

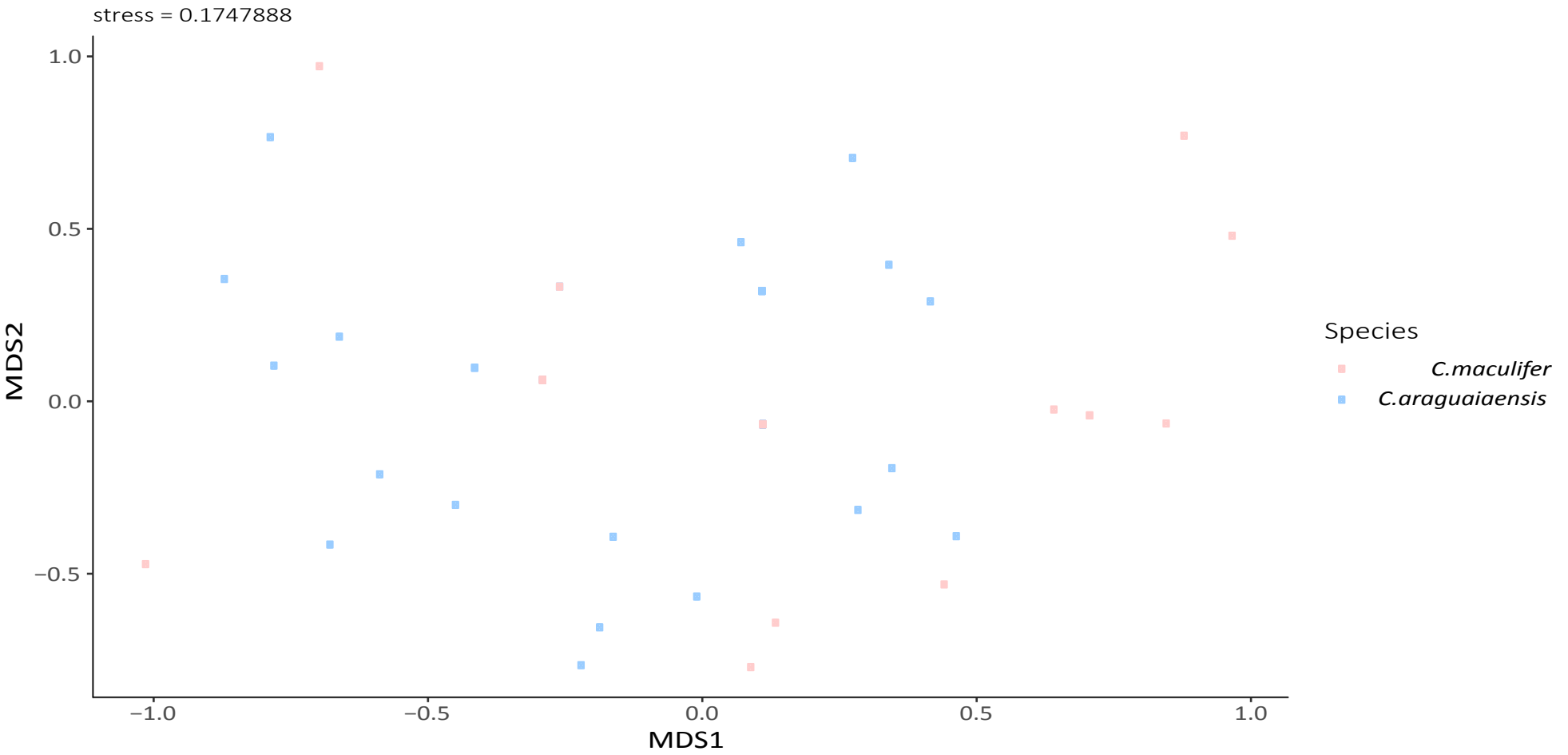


Figure 4.8: MDS visualisation of parasite communities and abundances across host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid). Distances calculated using Bray Curtis matrices.

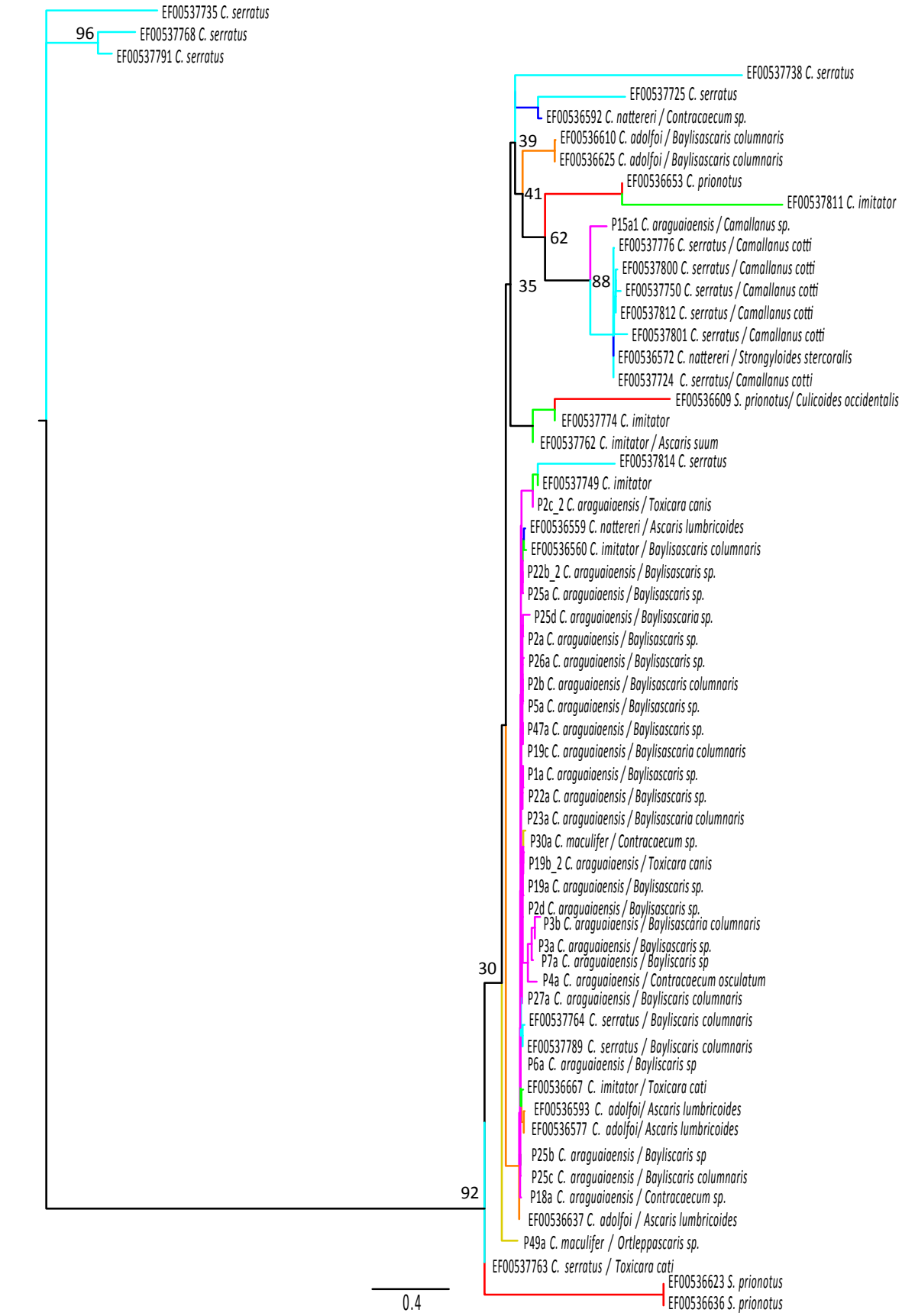


Figure 4.9: Topology recovered from the phylogenetic analysis of nematode CO1 genes. Trees were built in IQ-TREE utilising HKY+G4 model. Figures at nodes represent bootstrap support. Tree is coloured according to host species and labeled by sequence ID, host species and highest parasite blast hit (sequence ID host species / parasite blast hit).

4.3.3 Immune gene association analysis

Parasite and immune gene data were available for 14 *C. maculifer* and 35 *C. araguaiaensis* individuals. No correlation between Non-synonymous SNP counts at TLR1 and TLR2 loci and parasite counts was observed in *C. maculifer* or *C. araguaiaensis* (Figure 4.10) (Spearmans rank, $p > 0.05$). Covariance between effects of SNP count and host species on parasite counts was assessed using an ANCOVA. Across the data set non-synonymous SNP counts across pooled host species were found to have a significant effect (ANCOVA; $F = 4.89$, $df = 1$, $p < 0.05$) with greater SNP counts being negatively associated with parasite load, but host species and SNP count per host species did not.

C. maculifer had only a single SNP in TLR1 or TLR2, and as a result had negligible individual variation. Furthermore, haplotypes could not be separated out in the putative tetraploid *C. araguaiaensis* because sequences could not be fully phased along the full length of the TLR. As a result of these limitations to the data set investigations as to whether there was a relationship between parasite burden and TLR1/TLR2 characteristics, phylogenetic trees were constructed from consensus sequences (with degenerate bases included) of TLR1 and TLR2 in *C. araguaiaensis* and coloured by ranked parasite abundance (Figure 4.11). This analysis aimed to identify (without phasing) if individuals carrying specific TLR SNP profiles had greater of lesser parasite abundances. Parasite abundance did appear higher in individuals placed in the second cluster of TLR1 genes in *C. araguaiaensis* (Figure 4.11: tree A). Signals were less apparent in TLR2 but there were a higher proportion of infected hosts in the second cluster of TLR2 sequences (Figure 4.11: tree B).

Association analyses were performed on individual non-synonymous SNPs for both TLR1 and TLR2 in *C. araguaiaensis*. Overall none of the SNPs were significantly associated with parasite load. When p values from the association analysis were transformed (log10) and plotted in a Manhattan plot there were no clear peaks in either TLR1 or TLR2 suggesting little or no association with parasite burden (Figure 4.12).

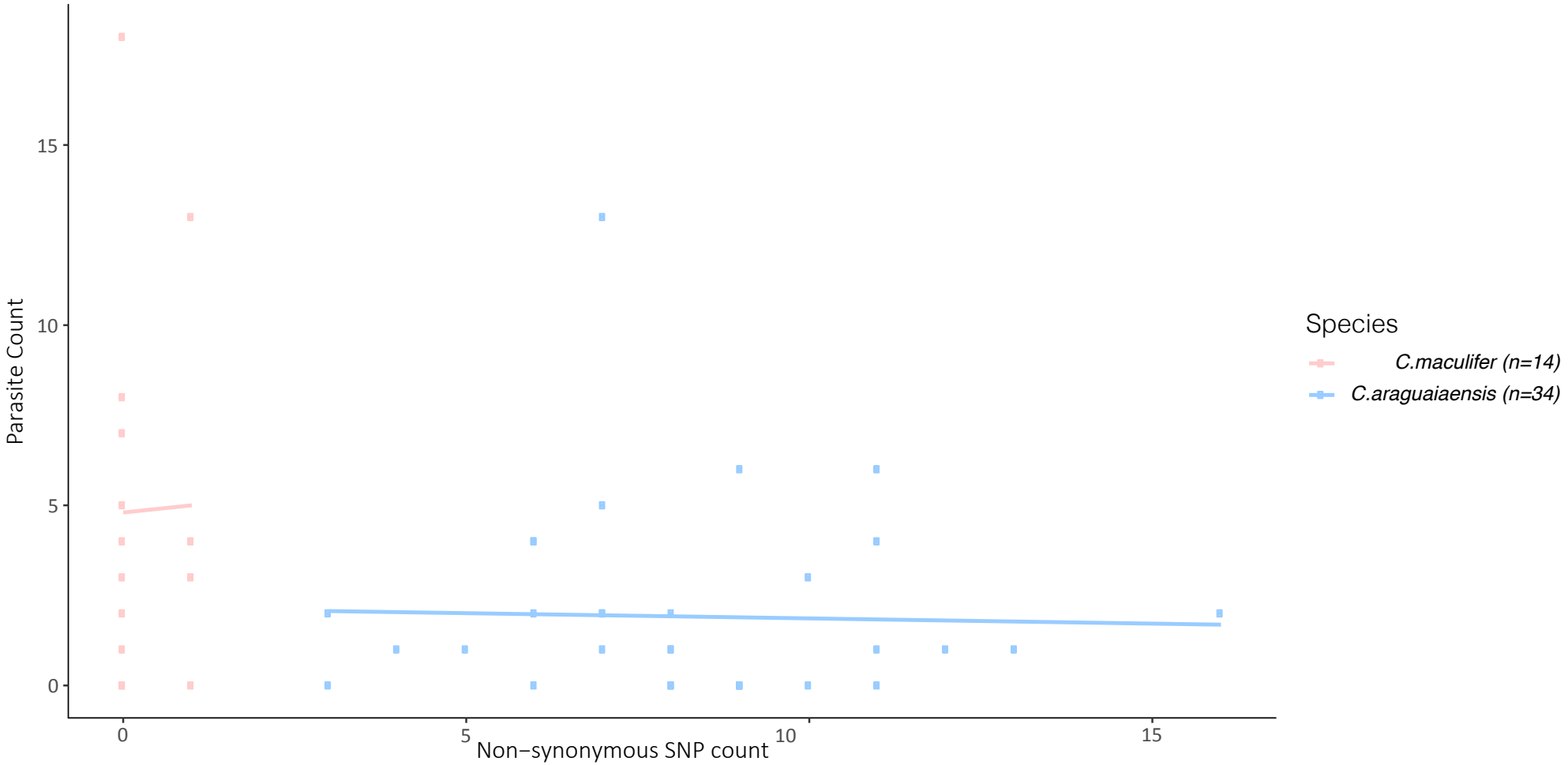


Figure 4.10: Non-synonymous SNP count in TLR1 and TLR2 of host Corydoras catfish species, C. maculifer (diploid) and C. araguaiaensis (putative tetraploid), plotted against overall parasite count. ANCOVA showed that SNP count across both species had a significant effect ($p<0.05$) but there was no significant effect of host species on parasite count.

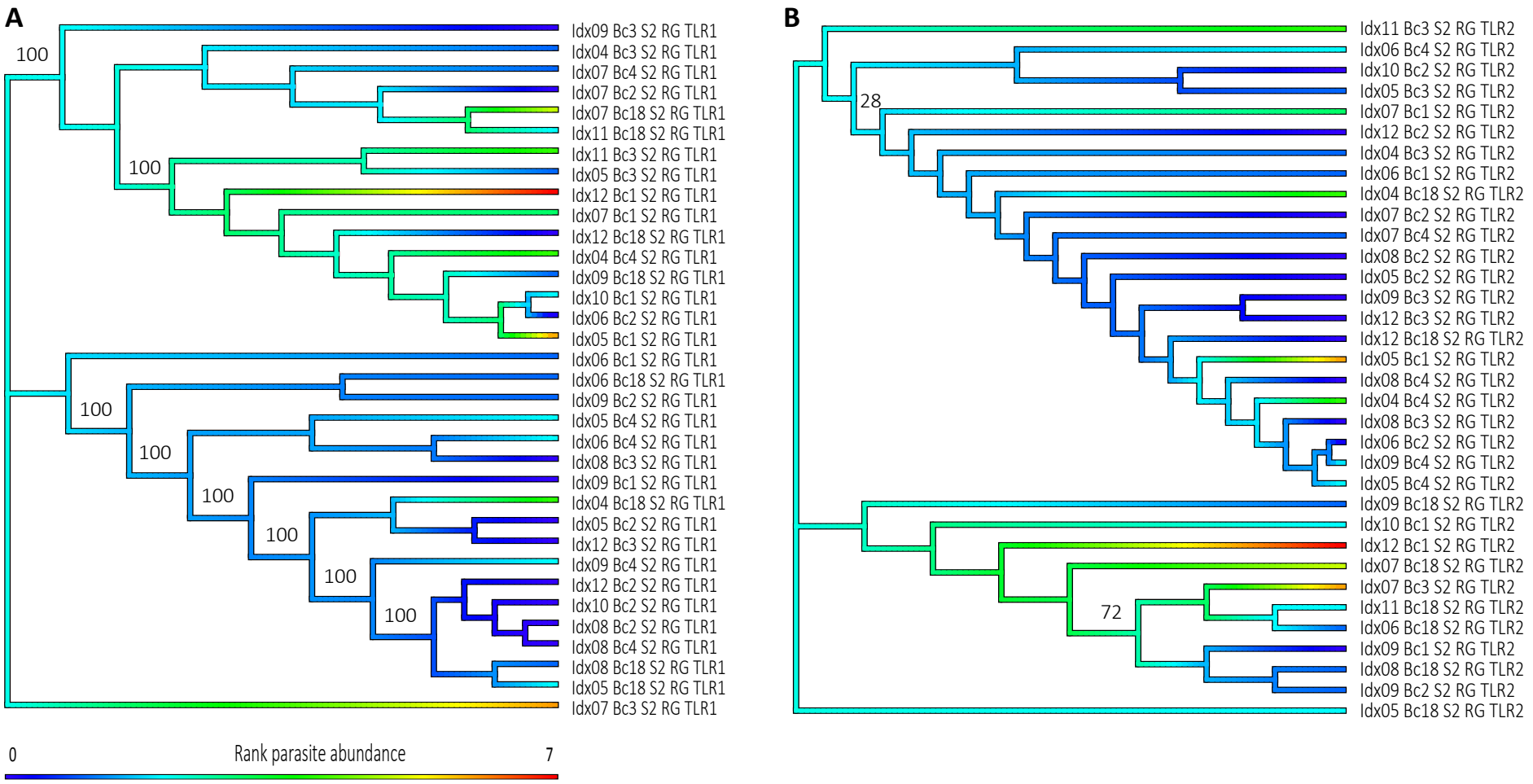


Figure 4.11: Tree topology recovered from maximum likelihood analysis for TLR1 (A) and TLR2 (B) in C. araguaiaensis, trees were built using the F81 model for TLR1 and the JC model for TLR2, converted to an ultrametric format and coloured according to ranked parasite intensity. Numbers at nodes represent bootstrap support.

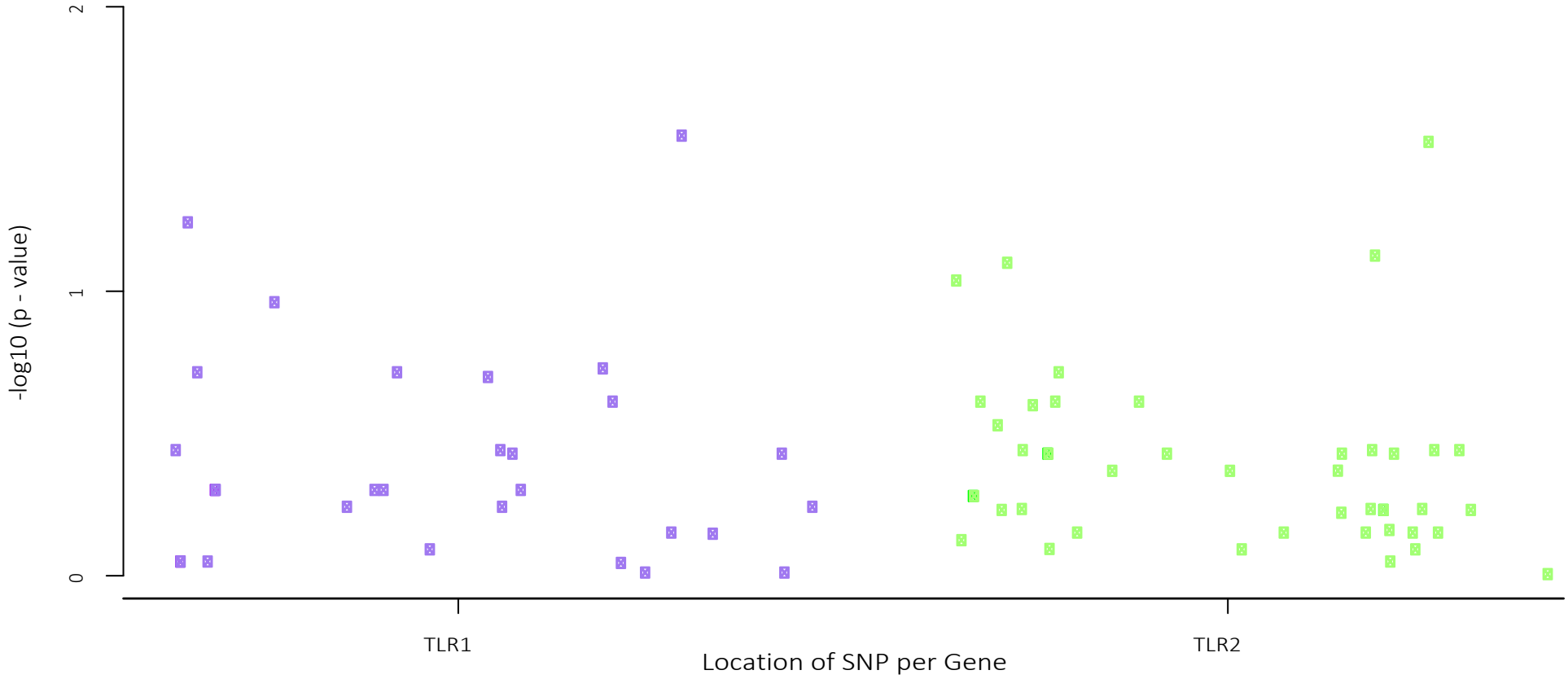


Figure 4.12: SNP association analysis for non-synonymous SNPs in TLR1 and TLR2 and parasite load in C. araguaiaensis. Coloured according to gene (i.e. TLR1 or TLR2). Association model qtscores produced using a Gaussian distribution.

4.4 Discussion

This chapter assessed the parasite burden of two sympatric coexisting species of *Corydoras* with differing ploidy states alongside previously assessed genetic diversity from two immune genes, TLR1 and TLR2. We previously ascertained (Chapter 3) that both TLR1 and TLR2 had significantly higher functional diversity in *C. araguaiaensis* (putative tetraploid) then in *C. maculifer* (diploid). We therefore examined parasite burdens between these two species to try and identify if there was a link between parasite load and immune gene diversity in these host populations.

When examining parasite count data, prevalence (i.e. the proportion of hosts infected) was largely similar between the two host *Corydoras* species (with the exception of external parasite prevalence which was significantly higher in *C. maculifer*). Conversely when looking at parasite intensities (i.e. the number of parasites per infected host individual) *C. maculifer* had greater parasite burdens then *C. araguaiaensis* in all tissues. This would suggest that *C. maculifer* and *C. araguaiaensis* have a similar likelihood of being exposed to parasites, but that there is a difference in the parasite tolerance and/or resistance between the two host species. The ability to limit the damage caused by parasitic infections is defined as host tolerance, while the capacity to limit overall parasite burden is host resistance (Råberg, Graham and Read, 2009). These data may indicate that *C. maculifer* is more tolerant of parasitic infections, or alternatively, that *C. araguaiaensis* is more resistant however this is impossible to resolve from these data. There were no significant differences detected in intensity between *C. maculifer* and *C. araguaiaensis* despite the evidence from Figure 4.2. A potential reason for this is illustrated when looking at parasite abundance (i.e. the number of parasites per host individual including uninfected individuals, Figure 4.3). The number of individuals with tissue specific infections is very low meaning that the intensity measure sample sizes (where zeros are removed) are also very low; this reduces the power of any statistical analysis and can produce misleading outcomes by indicating either falsely negative or positive results (Button *et al.*, 2013).

Other factors may influence parasite burden such as age, size, behaviour, physiology, population size, location and habitat (Ryce, Zale and MacConnell, 2004; Khan, 2012; Lester and McVinish, 2016). Samples were collected in the same stretch of river in the Araguaia river in Brazil and in some cases both species were caught in the same net. This ruled out potential influences of difference of location or habitat because both host species were deemed to share both. Positive correlations have been observed between fish length, age and parasite intensities in previous meta-analyses (Lo, Morand and Galzin, 1998; Poulin, 2000). This relationship was explored within these data (working on the assumption that size is correlated

with age). Positive correlations were observed between standard length and parasite abundance in both host species, although the correlation was only found to be significant in *C. araguaiaensis*. These data are difficult to compare between the host species. *C. maculifer* are generally larger then *C. araguaiaensis* and there is no overlap in size at the lower end of the standard length spectrum. As a result, a GLM was used to predict parasite abundance per mm of length per host species. This analysis indicated that even with size accounted for there were significant differences in parasite abundance both externally and across the total sum abundance between the two species with *C. araguaiaensis* individuals having lower parasite abundances than *C. maculifer* individuals. This supports the theory that *C. maculifer* is better able to tolerate higher parasite burdens (*C. araguaiaensis* that become equally heavily infected might die and therefore be removed from prospective samples), or it might be that *C. araguaiaensis* is more resistant to parasitic infections. Without performing population wide experiments under laboratory conditions, the tolerance verses resistance argument cannot be completely resolved.

Nutrition and dietary preference also play a role in parasite burden, with organisms at higher trophic levels generally harbouring greater parasitic abundances (Lester and McVinish, 2016). Previous stable isotope analysis suggested that *C. maculifer* operate at a lower trophic level then *C. araguaiaensis* (Alexandrou *et al.*, 2011). Therefore *C. araguaiaensis* should be more likely to harbour greater numbers of parasites. However the reverse is indicated by the data from the present study. It might be that the difference in trophic levels is not great enough to merit any substantial difference in parasite burden. Alternatively, this trend could be taken as an indication of greater parasite resistance in *C. araguaiaensis*.

If resistance were a factor this would stem from some form of immune advantage in C. *araguaiaensis*. Immune gene data is available from two TLR genes (TLR1 and TLR2) in both of these host species. When looking at the numbers of non-synonymous SNPs in both TLRs no correlations were observed in either species but given the coarseness of this metric and the lack of immune gene diversity in *C. maculifer* this is not surprising. There was more diversity in the immune genes in *C. araguaiaensis* and almost none in *C. maculifer*, so further analyses were restricted to this host species. Phylogenetic analysis of consensus sequences broke TLR1 into three major clusters; individuals belonging to the second cluster had a generally higher parasitic intensity then the other clusters. Meanwhile TLR2 was divided into two major clusters and individuals carrying genotypes from the second cluster again showed generally higher ranked parasite intensity. Without phasing out individual TLR haplotypes it is not possible to show if specific haplotypes are associated with higher or lower parasite burden. However this

analysis indicated that individuals within these clusters might potentially carry haplotypes associated with greater parasite intensity.

Individual haplotypes could not be identified in this study, however a broad SNP association analysis could be carried out. This test was limited by a low sample size and high respective SNP to sample size ratio, which is a limiting factor to the overall power of the analysis. Two small peaks associated SNPs with parasite abundance in both TLR1 and TLR2 however no individual SNP was statistically significantly associated with parasite abundance. This analysis was conducted across a high proportion of SNPs with a low overall sample size, which will have reduced the overall power of the analysis, it may be that different patterns would emerge if sample size was larger or SNP counts smaller. Previous studies have identified a number of associations between TLR polymorphisms and disease susceptibility, although the effects of these polymorphisms are normally deleterious (Noreen and Arshad, 2015). To investigate the impacts of specific immune gene polymorphisms and disease susceptibilities fully, experimental exposures would need to be carried out or large scale genetic association studies completed on this system, both of which are beyond the scope of this thesis.

Parasite community analysis indicated a general increase in encysted parasites in *C. maculifer* compared to *C. araguaiaensis* in all tissues except the liver, in which C. araguaiaensis showed higher intensities and abundances. Nematode parasite numbers were also higher in *C. maculifer* in all tissue types. However *C. araguaiaensis* was the only host to harbour isopod or acanthocaphala parasitic infections. MDS plots based on parasitic community data did not show any strong segregation of host species nor did phylogenetic analysis of the nematode sequence data.

Parasitic community abundances (Figure 4.7) again served to demonstrate the patchy nature of the parasitic data and help to explain why no strong community breakdowns could be established. It was rare for host individuals to share infections from the same parasitic taxa in the same tissue, which resulted in many data points being identified as outliers. To get a clearer view of parasitic communities between these host species, host sample sizes may need to be greatly increased. In addition to this the majority of parasitic identifications were only based on phylum as a higher degree of taxonomic resolution could not be established from morphological analysis and due to poor DNA quality/quantity many of the nematode samples tested using PCR failed to amplify.

The nematode sequence data were most closely aligned to known *Baylisascaris*, *Contracaecum*, *Toxocara*, *Camallanus* and *Ortleppascaris* species, with the majority being most closely associated with *Baylisascaris* according to GenBank. These are all types of parasitic roundworm. *Baylisascaris* and *Toxocara* species are most commonly associated with

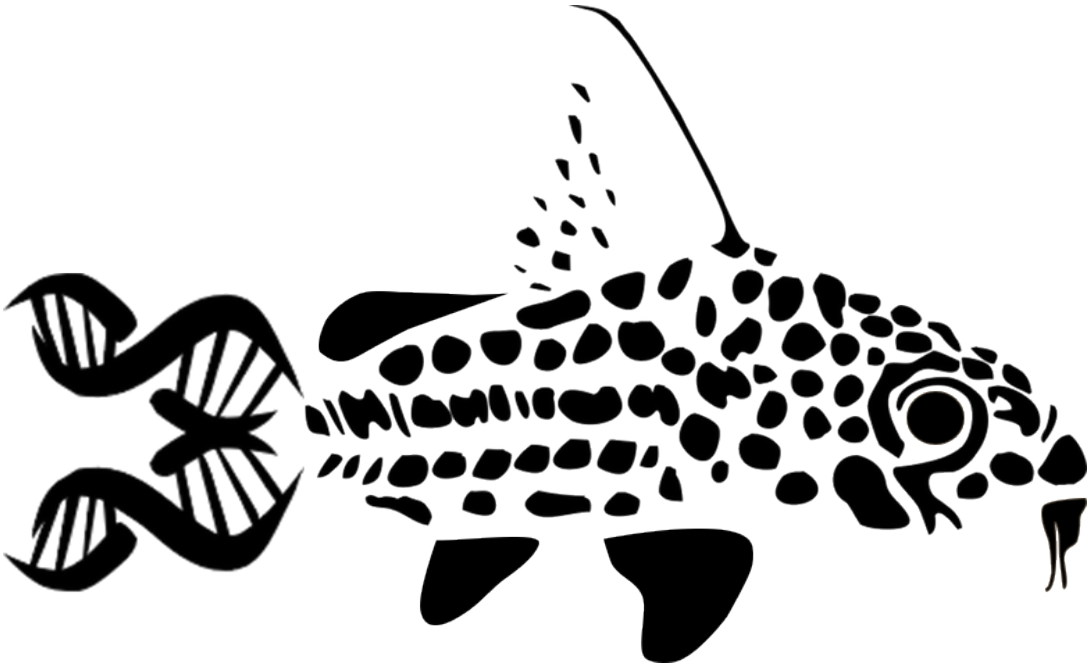
mammalian hosts, the most well-known species being *Baylisascaris procyonis* (the raccoon roundworm) and *Toxocara canis* or *T. cati* (the dog and cat roundworms) (Strube, Heuer and Janecek, 2013; Sapp *et al.*, 2017). *Contracaecum* species have been associated with fish, cephalopods, marine mammals and fish eating birds (Szostakowska, Myjak and Kur, 2002). *Camallanus* are generally associated with fish hosts although some species also infect amphibians and reptiles (Kuzmin *et al.*, 2011) and *Ortleppascaris* is associated with amphibian hosts (Pereira e Silva, Furtado and Nascimento dos Santos, 2014). The general association of the majority of these parasites for non-fish hosts suggests that these *Corydoras* hosts may not have been the definitive host for these parasites but might have acted in an intermediary capacity. *Corydoras* are benthic omnivorous detritivores (Nijssen, 1970) and may ingest mammalian faecal matter containing parasites or their eggs. However, information on potential predators that might prey on the *Corydoras* is very sparse. It might be that the *Corydoras* act as an intermediary host facilitating the transmission cycle of the parasites they carry, or they might be accidental hosts and constitute a dead end in the parasite transmission cycle. More information on the life histories of both host and parasite would be required to further explore this relationship. It is also worth noting that, although these species identifications were based on similarity to species already recorded in GenBank, species of a similar identification frequently failed to cluster in Figure 4.9. This sheds a degree of doubt on the initial species identifications as registered in GenBank especially given the intrinsic difficulties associated with morphological ID of nematodes.

4.4.1 Conclusion

This chapter has examined the parasite prevalence's, intensities and abundances of both count and community data in two species of sympatric coexisting species of *Corydoras* catfishes (*C. maculifer* and *C. araguaiaensis*) with differing genome sizes. Both species had roughly equal parasite prevalence's (number of infected individuals) however *C. maculifer* (diploid) was found to generally have greater parasitic intensities (numbers of parasites per infected individual) than *C. araguaiaensis* (putative tetraploid). This suggests that either *C. maculifer* has higher parasitic tolerance or that *C. araguaiaensis* has higher resistance to parasites. Immune gene analysis showed that some TLRs diversity might be associated with higher or lower parasite burden but without haplotype phasing further analysis was not possible. SNP association analysis found no significant links between specific SNPs and parasite burden, however the high SNP to sample ratio means that the overall power of this analysis is low. Community analyses also provided little resolution between the two species due to sample size limitations. However blast analysis of nematode sequence data indicated that many of the

nematode parasites belong to genera that are often associated with non-fish definitive hosts. This might indicate that these host species are more commonly intermediary hosts as opposed to definitive hosts. Overall this chapter has highlighted the beginnings of some potentially interesting trends. Further assessments over larger host sample sizes, greater understanding of both host and parasite life histories and analysis across a greater immune gene repertoire would shed further light on the questions this chapter has raised.

Chapter 5:
Characterisation of pathogen recognition receptors in *Corydoras maculifer*



5.1 Introduction

The principal function of the animal immune system is to detect and respond to invading pathogens. In order to mount a successful immune response a host organism must first be able to detect foreign pathogenic antigens (Takeda and Akira, 2005). The pathogen recognition receptor (PRR) gene family is partially responsible for this (Rajendran *et al.*, 2012). The PRRs that these genes encode are evolutionary conserved, germline-encoded proteins with broad specificities and are a vital part of the innate immune system (Alvarez-Pellitero, 2008; Meng *et al.*, 2009). They recognise conserved pathogen associated molecular patterns (PAMPS) and initiate rapid immune responses (Alvarez-Pellitero, 2008; Meng *et al.*, 2009). Three families of PRR have been identified; the toll like receptors (TLRs), nucleotide-binding oligomerization domain (NOD) and leucine rich repeat containing receptors (NLRs) and retinoic acid inducible gene 1 (RIG-1) like helicases (RHLs) (Aoki and Hirono, 2006; Chang *et al.*, 2011; Rajendran *et al.*, 2012).

The TLR gene family encode a group of type I transmembrane proteins that recognise pathogen associated molecular patterns (PAMPs) and initiate downstream immune mechanisms (Zhao *et al.*, 2013). Their basic shared structure is composed of an N-terminal ectodomain, multiple leucine rich repeats (LRRs), a C-terminal domain, a transmembrane region and a toll-interleukin receptor (TIR) signalling domain (Medvedev, 2013). In the channel catfish, *I. punctatus*, fifteen TLRs have been identified (including TLR1, TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, TLR9, TLR18, TLR19, TLR20, TLR21, TLR22, TLR25, TLR26). Subfamilies of TLRs are also associated, broadly, with different pathogenic vectors, TLR1, TLR2, TLR4, TLR5 and TLR9 have previously been linked with bacterial infections while TLR3, TLR7, TLR22 and TLR8 have been associated with viral infections (Pietretti and Wiegertjes, 2014). Once activated TLRs have a direct role in the initiation of the innate inflammatory response and an influential role in regulation of antigen presentation on dendritic cells (Salaun, Romero and Lebecque, 2007).

The NLR protein family are usually characterised by three distinct domains; an N terminal protein interaction domain, such as the caspase recruitment and activation domain (CARD) or pyrin domain (PYD), a nucleotide binding domain (NACHT) and a C-terminal LRR domain (Meng *et al.*, 2009; Rajendran *et al.*, 2012). These proteins tend to be intracellular PRRS and have a role in inducing the inflammatory response and/or apoptosis (Rajendran *et al.*, 2012). NLR regions are thought to be associated with different functions, with the C-terminal region being linked to potential PAMP recognition, the NACHT domain being a self-regulatory region and the N thermal domain being thought to have involvement in protein-protein interactions, signal transduction and initiation of downstream immune reactions

(Rajendran *et al.*, 2012). A total of 22 NLRs have been identified in *I. punctatus*, including; NOD1, NOD2, NOD3a, NOD3b, NOD4, NOD5, NLR-B1, NLR-B2, NLR-C1 to NLR-C11, Apaf1, CIITA and NACHT-P1) (Rajendran *et al.*, 2012).

The final PRR gene group are the RHLs, these act as cytocolic receptors belonging to the DExD/H box RNA helicases and have roles in the recognition of viral RNA (Liu *et al.*, 2016). Three classes of RHL have been identified in fish; RIG-I has been identified in a number of Cypriniformes, Siluriformes and Salmoniformes species whereas MDA5 and LGP2 have been found more generally across the Acanthopterygii (Chen, Zou and Nie, 2017). It has not yet been confirmed if RIG-I has been lost in some Acanthopterygii or if it simply hasn't been successfully isolated yet (Chen, Zou and Nie, 2017).

The Corydoradinae are a species rich subfamily of freshwater Neotropical catfishes found across South America (Fuller and Evers, 2005). *Corydoras maculifer* is a diploid species within the Corydoradinae, found in the Araguaia region of Brazil. Earlier investigations found very little variation in two classes of PRR (TLR1 and TLR2) within this species (Chapter 3). Across a population of seventeen individuals one non-synonymous single nucleotide polymorphism (SNP) was identified in TLR1 and one synonymous SNP in TLR2. This finding was considered exceptionally low for immune gene diversity, with similar SNP counts being identified in populations that had recently been through a bottleneck or were in threatened populations (Grueber *et al.*, 2015; Gilroy *et al.*, 2017).

5.1.1 Aims and objectives

The immune gene diversity *of C. maculifer* in TLR1 and TLR2 was found to be very low in population wide analyses (Chapter 3). As a result this chapter aimed to identify and characterise other PRRs from TLR, NLR and RHL gene families across the *C. maculifer* genome and identify if any further diversity is apparent in any of these other genes and gene families in a single individual. Genomic and transcriptomic sequencing data were available for *C. maculifer* so we aimed to explore PRRs using genome annotation, mining and mapping techniques to address the question of PRR diversity in *C. maculifer*.

5.2 Methods

5.2.1 Sampling and DNA extraction

One wild caught individual *C. maculifer* was collected from the Araguaia region in Brazil by MIT, CO and EB in 2015, euthanized by anaesthetic overdose and stored in 100% ethanol. DNA was extracted from fin clip tissue using a Qiagen DNeasy Blood and Tissue extraction kit. DNA concentration was quantified using a nanodrop8000 spectrophotometer (Thermo Scientific) and fragment size measured via gel electrophoresis on a 1.2% agarose gel spiked with ethidium bromide prior to sequencing.

5.2.3 Sequencing, Assembly and Scaffolding

Genome sequencing was carried out on the *C. maculifer* genomic DNA. Library preparation using PCR-free protocols for adaptor ligation and sequencing on a range of platforms was carried out by the Earlham Institute, Norwich Research Park, Norwich. One library was sequenced on a single Illumina HiSeq2500 lane using a 250bp paired-end read metric. Following this, twelve Nextera long mate pair (LMP) libraries were constructed, with different size fractions, from a single genomic DNA sample. These LMP libraries were sequenced on a single Illumina MISeq lane with a 300bp paired-end read metric. Two of these libraries were selected on the grounds of having the largest insert size (named LIB21508 and LIB21509 with average insert sizes of 8678.2bp and 8730.0bp respectively) and were then sequenced on a lane of Illumina Hiseq2500 with a 250bp paired end read metric.

Paired-end PCR-free libraries were assembled using w2rap-contigger (Clavijo *et al.*, 2017) under default settings to produce a set of contiguous units (contigs). Long mate-pair libraries were cleaned using NextClip and SOAPdenovo2 was used under default setting to combine contigs and mate-pair libraries into scaffolds. The scaffolding pipeline was tested using multiple kmer sizes (kmer sizes = 17, 19, 21, 25, 31 and 49) and finally run with a kmer size of 19 because this rendered the highest N50 and N90 scores. Assembly and scaffolding processes were carried out by Sarah Marburger and Levi Yant of the John Innes Centre, Norwich Research Park, Norwich.

Transcriptome sequencing was undertaken in collaboration with Professor Claudio Oliveira (Botucatu-Unesp) following total RNA extraction from a number of *Corydoras* skin tissue samples (*C. aeneus* x 2, *Aspidoras fuscoguttatus* x 2, *C. haraldschultzi* x 2, *C. schwartzi* x 1, *C. elegans* x 1, *C. fowleri* x 1, *Scleromystax prionotos* x 3*, C. paleatus* x 1, *C. melini* x 1, *A. pauciradiatus* x 1, *C. julii* x 2, *C. nattereri* x 3, *C. araguaiaensis* x 3, *C. maculifer* x 2, *C. polystictus* x 1, *C. hastatus* x 2, *Aspidoras sp.* x1, un-described lineage 1 x 1, un-described

lineage 8 x 1, un-described lineage 9 x 1). RNA was sequenced as a paired-end metric on two Illumina HiSeq lanes in the Laboratorio de Biotechnologia, Brazil. Libraries were de-multiplexed and cleaned using Trimmomatic (version 0.2.36) Paired end reads were individually assembled by library using default settings on the Trinity DeNovo assembler (version 2.6.9, Grabherr et al. 2011) that was run through the online Galaxy server. Transcriptome clean-up and assembly was run by Ellen Bell.

5.2.4 Quality checks and Annotation

Benchmarking Universal Single Copy Orthologs (BUSCO version 3.0.0, Waterhouse et al. 2018) was used to assess the assembly and completeness of the *C. maculifer* genome. This program uses a database of BUSCOs, which are expected to be in all genomes as a single copy, to determine the completeness and level of assembly derived fragmentation in a genome (Simao *et al.*, 2015). BUSCO was run using the Actinopterygii database (actinoptergii-obd9 created 13-02-2016) and denoting zebrafish (*Danio rerio*) as its closest existing species in the Augustus documentation (a parameter set to assist with gene finding).

To annotate the *C. maculifer* genome the Genome Sequence Annotation Server (GenSAS, version 5.0, Humann et al. 2018) was used. This server acts as a pipeline to find and mask repetitive elements (via RepeatMasker and RepeatModeler programs), predict locations and identities of genes (using a combination of Augustus, FgeneSH, Genscan, Glimmer3, GlimmerM, SNAP, tRNAScan, getorf, BLAT and BLAST, EvidenceModeller programs) and then visualises and curates its output (via WebApollo and Jbrowse) before publishing it as GFF2 and FASTA files. For annotation of the *C. maculifer* genome, assembled scaffolds were uploaded into the GenSAS server and a concatenated file of assembled transcriptome sequence data from multiple *Corydoras* species was provided as evidence for gene prediction. Gene prediction software accumulates evidence of gene presence using a range of data sources including RNA sequence data. RNA sequence data may be aligned to genome scaffolds and acts as evidence that potential genes are expressed and are therefore present (Yandell and Ence, 2012). Quality checks and annotation of the *C. maculifer* genome was completed by Ellen Bell and Martin Taylor.

5.2.5 Immune gene mining and analysis

A list of Fish PRR associated genes was constructed following a literature search. This list included all TLRs (TLR1, TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, TLR9, TLR18, TLR19, TLR20, TLR21, TLR22, TLR25, TLR26), NLRs (NOD1, NOD2, NODC3, NODC5, NLRX1) and RIG (RIG1) associated

genes and was used to search within the annotated *C. maculifer* genome. Using the command line programme grep, it was possible to identify matches in gene IDs from the gene list and the GFF annotation files. This search gave positional information for potential hits, which were subsequently extracted from the assembled *C. maculifer* genome and trimmed to only include coding regions. In order to identify any genes that might have been missed during annotation, a fasta file was also assembled from all known TLRs, NODs, NLRs and RIG genes from *Ictalurus punctatus* and *Danio rerio* (sequences downloaded from the National Centre for Biotechnology Information (NCBI) Genbank). The assembled *C. maculifer* genome was then converted into a BLAST database and the PRR gene list was blasted against it under default settings (Blastn, NCBI-2.2.29). Blast hits were filtered to include only those with greater then 80% similarity and over 500bps in length before being isolated from the assembled genome and trimmed as before.

All genome-extracted sequences were loaded into Geneious (version 9.1.8) and the inbuilt Open Reading Frame (ORF) finder was used to put all of these sequences in frame and translate them. All sequences were then blasted against the NCBI Genbank to check their ID and then passed to the Single Modular Architecture Research Tool (SMART, Letunic & Bork 2018) to check that they contained domains expected for genes of their class.

Once confidence in the identity of the genome extracted PRR sequences had been established they were used as reference sequences to map raw *C. maculifer* genome reads back to. Raw genome reads were interleaved using an in-house Python script and mapped back to the reference sequences using BWA-mem (version 0.7.12, Li & Durbin 2009). This process was run with the following stringency parameters, a mismatch penalty of 20, a gap open penalty of 30 a gap extension penalty of 10 and a clipping penalty of 50. Mapped reads were further filtered to only include reads with a mapping quality score of 30 or greater and hard and soft clipped reads were removed from the final output. SNPs and haplotypes were then called using QualitySNPng (Nijveen *et al.*, 2013) which was configured to require a minimum number of five read counts per allele and a minimum of 10% overall read depth per allele.

Amino acid sequences for TLRs, NODs and NLRs from translated regions of the *C. maculifer* genome were aligned to published *Ictalurus punctatus* and *Danio rerio* sequences in Geneious. Trees were built from aligned sequences using IQ-TREE (version 1.5.5, Nguyen et al. 2015; Hoang et al. 2018), which built maximum likelihood trees based on the best model fit identified by jModelTest using the Bayesian information criterion (BIC). Trees were built using 1000 ultrafast bootstrap replicates and visualised in FigTree (version 1.4.3).

5.3 Results

The *C. maculifer* genome was assembled into 503668 contiguous units with an N50 of 30744bp these were then brought together into 1948 scaffolds with an N50 of 607365. Summary statistics for contig and Scaffold assembly of the *C. maculifer* genome are displayed in Table 5.1, along with the outputs from the BUSCO analysis. BUSCO analysis suggested that the *C. maculifer* genome contained 88.2% of its expected content in an un-fragmented form (Table 5.1).

Genome annotation and Blast searches found a set of 13 potential TLR, NOD or NLR associated hits, including; TLR1, TLR2, TLR7, TLR18, TLR25, NOD1, NOD2 and six NLRP3 associated genes (Table 5.2). No RIG like receptors were identified. Mapped reads were exceptionally high for TLR18, but further examination showed that the majority of these reads were from a small region at the 5' end of the gene and of low quality.

Variant SNP calling across the TLR and NOD families detected a single SNP in TLR1 which was identical to the single SNP identified in TLR1 in the earlier population-wide TLR analysis (Chapter 3). Four SNPs were identified in TLR18, but as these all clustered across the poorly mapped region, sequencing error was likely to be high and consequently they were excluded from this analysis. Similarly, with the exception of TLR1 for which two haplotypes were identified, the remaining TLR and NOD genes were present as a single haplotype. The TLR genes all clustered with their TLR counterparts in *Ictalurus punctatus* and *Danio rerio* following phylogenetic analysis (Figure 5.1) and all showed signs of a similar protein domain structure with SMART analysis identifying multiple leucine rich repeats, a leucine rich repeat C terminal domain and a toll interleukine receptor region for each proposed TLR gene (Figure 5.2). Phylogenetic analysis showed the *C. maculifer* NOD1 gene clustering with its counterpart from *I. punctatus* and *D. rerio* but NOD2 was more distantly separated only being distantly grouped with the whole NLR cluster (Figure 5.1). Following SMART analysis a single caspase activation and recruitment domain (CARD) was identified in NOD2 and no significant architecture was identified for NOD1 by SMART analysis at all (Figure 5.2).

An additional set of six NLRP3 associated genes were identified from across the genome. These NLRP3s had varying numbers of SNPs that ranged from 0 to 60 and haplotype counts ranging from 1 to 9. All of these NLRP3 sequences were labelled as such within the *C. maculifer* genome annotation. NLRPs have not been found in *I. punctatus* of *D. rerio*. However NLRP associated genes from *C. maculifer* clustered with NLRC3, NACHT or NOD3 genes from *I. punctatus* or *D. rerio* following phylogenetic analysis, suggesting that differences in nomenclature are greater than the phylogenetic differences in the genes. The NLRP3 variant from scaffold 1940 clustered most closely with a sequence containing NACHT, LRR and PYD

domains in *I. punctatus*, while NLRP3 variants from scaffolds 737, 1928, 915 and 1630 clustered with a NOD3 receptor also identified in *I. punctatus*. NLRP3 on scaffold 590 only loosely clustered with the NLRs as a whole. All NLRP3 sequences were found to contain a fish specific NACHT associated domain (FISNA) and this was the only molecular architecture identified by the SMART analysis.

Table 5.1: Assembly, scaffolding and BUSCO summary statistics from the C. maculifer genome, demonstrating expected coverage and completeness.

Contig Assembly Summary Statistics				
Number of Contigs		**n:500 (number of contigs over 500bp long)**		**N50**
503668		108959		30744
Scaffold Assembly Summary Statistics				
Kmer size	**Scaffold number**	**Average scaffold length**	**N50**	**% of estimated genome size**
17	2185	282669	563517	97.16
19	1948	324890	607365	99.56
21	2031	312626	598544	99.88
25	2338	270932	476882	99.64
31	3400	186125	374957	99.54
49	5094	119036	221025	95.38
BUSCO Summary Statistics				
	Complete BUSCOs	**Fragmented BUSCOs**	**Missing BUSCOs**	**Total BUSCO groups searched**
Raw Count	4042	179	363	4584
Percentage	88.2%	3.9%	7.9%	

Table 5.2: Summary mapping statistics for TLRs, NODs and NLRs in the C. maculifer genome along with predicted SNP and haplotype counts from QualitySNPng.

Gene	Gene length	Number of reads mapped	Average read depth	Read depth standard deviation	Predicted SNP count	Predicted haplotype count
TLR1	2,820	1,425	126.9	33.2	1	2
TLR2	2,261	741	82.3	24.6	0	1
TLR7	4,802	1,880	98.3	23.5	0	1
TLR18	4,868	28,650	549.2	1718.9	0	1
TLR25	3,812	1,592	104.8	26.0	0	1
NOD1	766	271	88.6	42.4	0	1
NOD2	3,957	1,966	124.7	41.6	0	1
NLRP3, Scaffold 590	1,852	854	115.8	46.3	33	4
NLRP3/NACHT/NLRC3, Scaffold 1940	1,427	1,599	282.1	176.8	60	9
NLRP3/NOD3, Scaffold 737	1,853	892	120	37.3	19	3
NLRP3/NOD3, Scaffold 1928	1,861	416	56.0	20.5	0	1
NLRP3/NOD3, Scaffold 915	1,857	797	107.6	33.6	0	1
NLRP3/NOD3, Scaffold 1630	1,860	890	120.2	37.2	19	3

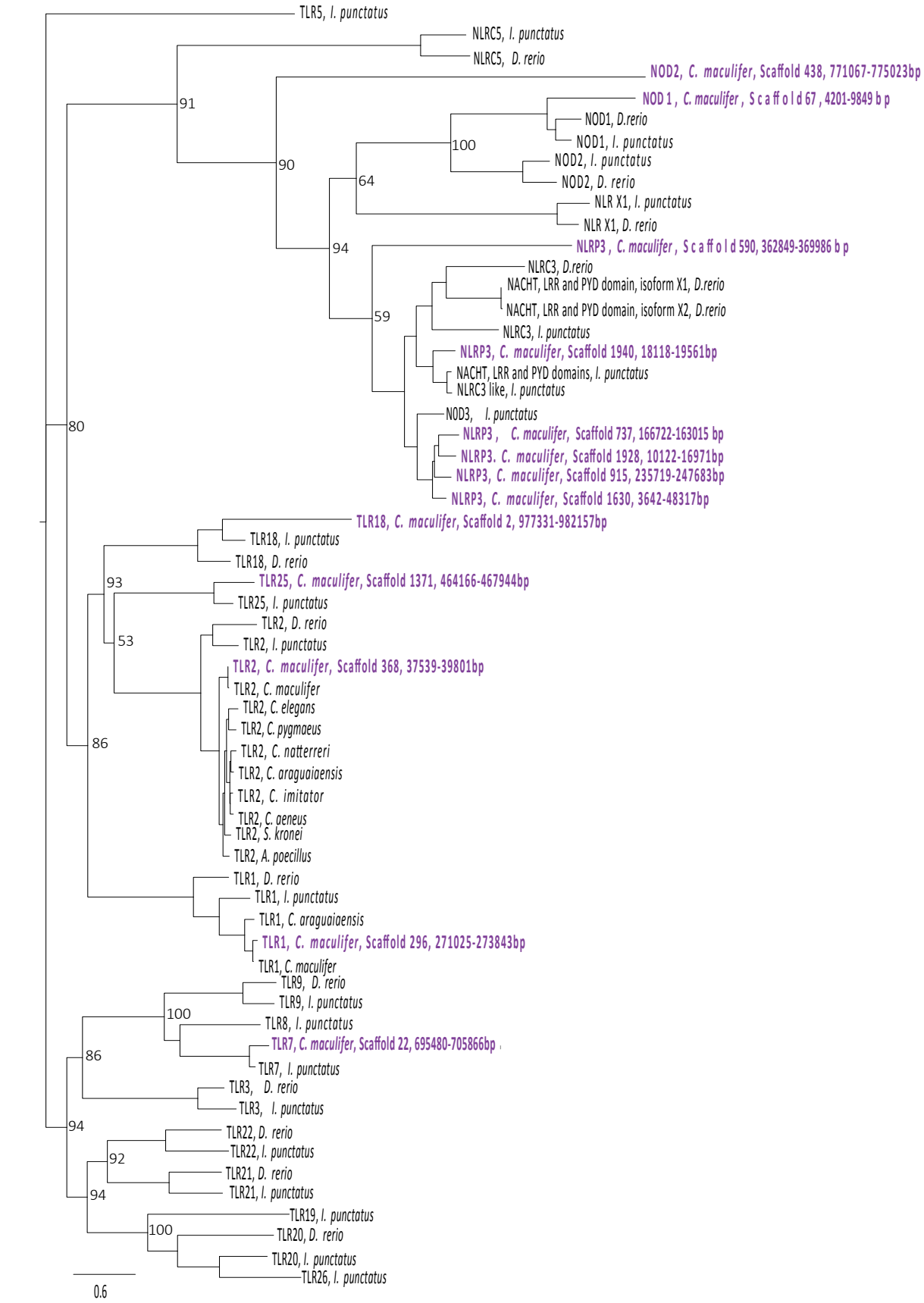


Figure 5.1: Topology recovered from the phylogenetic analysis of known TLR, NOD and NLR associated genes across the C. maculifer genome (purple), available Corydoras species data, I. punctatus (channel catfish) and D. rerio (zebrafish). Trees were built in IQ-TREE utilising VT+F+G4 model; figures at nodes represent bootstrap support.

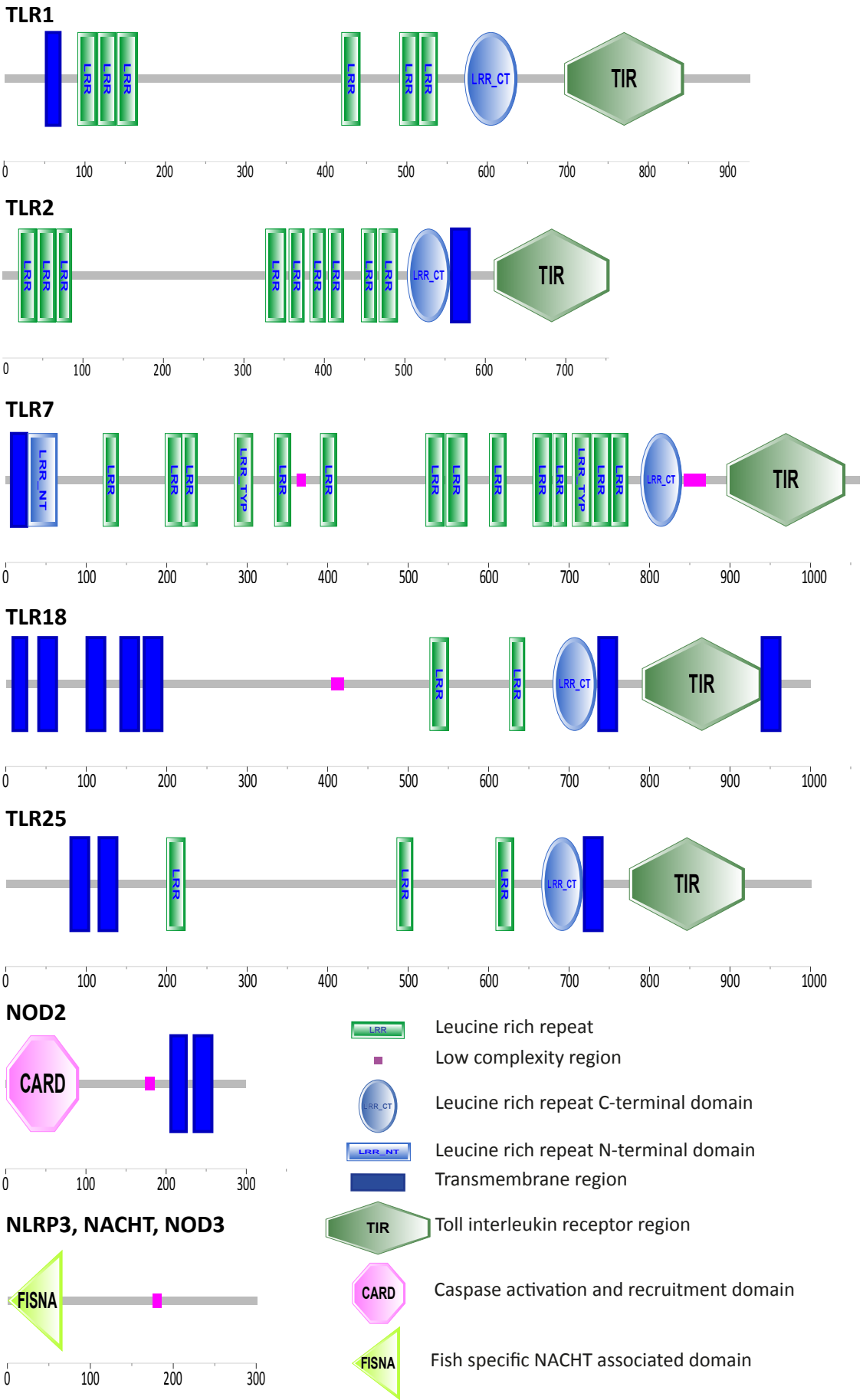


Figure 5.2: Protein domain prediction for TLR, NOD and NLRs based on SMART analysis.

5.4 Discussion

In this chapter two PRR gene families - TLRs and the NLRs - were identified and partially characterised from across the *C. maculifer* genome. Attempts were made to isolate RHL genes but none could be found in the existing *C. maculifer* genome. This may be because they were missed during the sequencing process; alternatively they might have been fragmented and lost during the various assembly and scaffolding processes.

Variation across the five TLR genes identified (TLR1, TLR2, TLR7, TLR18 and TLR25) was comparable to variation detected in Chapter 3 where a single SNP was identified in TLR1. Once again the only TLR gene to have reliable evidence of a SNP was TLR1, this SNP was in the same position as that identified in Chapter 3 and was non-synonymous. Variation was similarly low in NOD1, NOD2 and two of the NLRP3 genes (scaffolds 1928 and 915) with no SNPs identified in either. This supports previous observations in *C. maculifer* and is comparable with immune gene variation in populations that have been through a bottleneck (Grueber *et al.*, 2015; Gilroy *et al.*, 2017).

Greater genetic diversity was identified at the remaining NRLP3 genes, with SNP and haplotype counts ranging from 19 to 60 and 2 to 9 respectively. When assessing the phylogenetic and structural evidence for the diversity in these genes, a number of potential reasons for this increased diversity were identified. The NLRP3 genes from scaffolds 737, 1928, 915 and 1630 all cluster tightly and are most closely associated with NOD3 in *I. punctatus*. This could be evidence of a tandem duplication event across this gene family and if these genes are very similar then that can confound assembly processes (Bailey *et al.*, 2004). With tandemly duplicated genes an assembler aims to correctly place almost identical genes in multiple positions within the overall assembly, which can lead to either under or over representation of a duplicated sequence within the assembly (Bailey *et al.*, 2004). All evidence to date suggests that *C. maculifer* is diploid, however the range of SNPs detected in these sequences have led to high haplotype predictions for these genes. This would suggest that the NLRP3 cluster, which current analysis suggests is broken into six genes, might actually be composed of a greater number of genes, which are either structurally similar or tandemly duplicated. Tandem duplications have been implicated in the large NACHT domain repertoire observed in the coral *Acropora digitifera* so this observation in *C. maculifer* is not novel to the NLR gene family (Hamada *et al.*, 2012). Long-range amplicon sequencing with haplotype phasing might shed further light on this.

Analysis of the molecular architecture of these PRRs showed that TLRs had a relatively conserved structure, which conformed to that observed in the literature (Meng *et al.*, 2009; Rajendran *et al.*, 2012) with multiple leucine rich repeats, a C-terminal domain and a toll

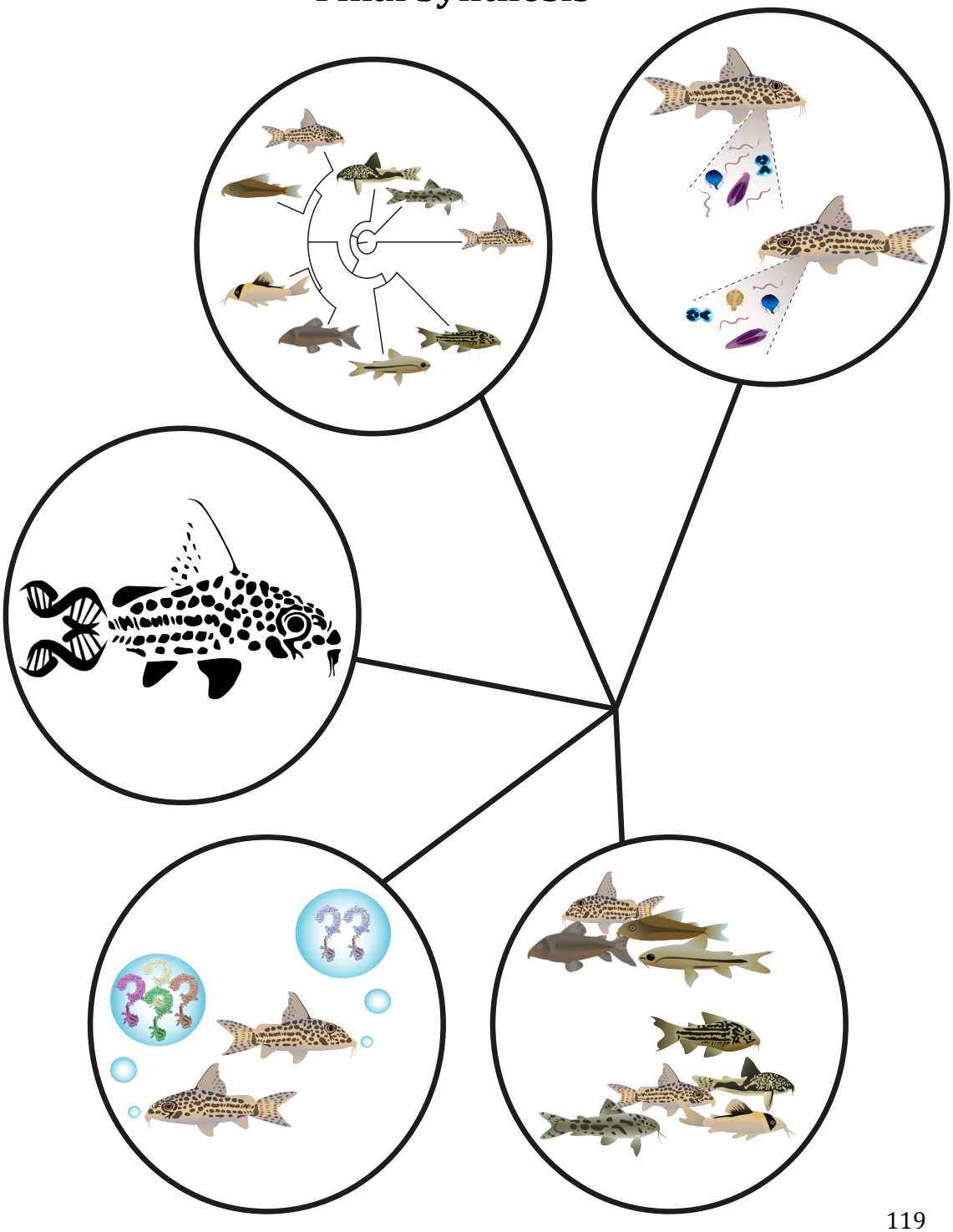
interleukin receptor region. An N-terminal domain was identified in TLR7 and numerous transmembrane domains were identified across the TLR family. Classically, transmembrane regions in TLRs are found beside the C-terminal region although some chicken (*Gallus gallus domesticus*) TLR variants have been shown to have transmembrane regions near the N-terminus (Temperley *et al.*, 2008). However SMART identified transmembrane regions by looking for hydrophobic signatures, it is therefore possible that SMART falsely identified some of these transmembrane regions (Temperley *et al.*, 2008).

No significant structural signatures could be detected by SMART for NOD1 and a CARD region was the only architecture detected in NOD2. A fish specific NACHT associated domain (FISNA) was detected in all NLRP3 sequences but no further architecture. A full NLR is generally identified by the presence of a CARD or PYD region, a NACHT/FISNA region and a C-terminal LRR domain (Meng *et al.*, 2009; Rajendran *et al.*, 2012). The presence of some, but not all of these domains across NODs and NLRP3s, suggests that although these sequences are similar to those found in other species they are not complete. These sequences could be truncated (Rajendran *et al.*, 2012) or they could be partially complete as a result of incomplete sequencing or assembly. Once again long-range amplicon sequencing would be an effective method for getting higher degrees of resolution over the presence, structure and function of the NLR gene family.

5.4.1. Conclusion

This chapter has identified five apparently complete TLR genes and eight potentially partially complete NLRs across the *C. maculifer* genome. The majority of these genes supported the low diversity observed in Chapter 3. Originally it was hoped that diversity in these PRRs could be compared across *C. maculifer* and *C. araguaiaensis* genomes. However genome coverage was very low in *C. araguaiaensis,* to the extent that very little read depth was available which meant that mapping was very uneven and a poor representation of the full gene length. This data was considered unusable without further sequencing efforts, however it would make an interesting comparison in future work.

Chapter 6:
Final Synthesis



6.1 Synopsis

The aim of this thesis has been to explore potential relationships between genome duplication, immune gene diversity and parasite load in the Corydoradinae catfishes. The research has examined elements of immune gene diversity at a range of different levels. It has assessed differences in an immune gene across representatives of each of the nine Corydoradinae lineages, observed how these differences manifest across two sympatric populations of *Corydoras* and subsequently assessed the parasite communities from this same *Corydoras* host community, and it has partially characterised a suite of immune genes across a single *Corydoras* genome.

6.1.1 Characterising TLR2 across the nine lineages

The Corydoradinae subfamily can be divided into nine distinct evolutionary lineages (Alexandrou *et al.*, 2011; Marburger *et al.*, 2018). Evidence suggests that this subfamily has undergone two whole genome duplication events, the first of which is thought to have occurred at the base of lineage 2 and the second encompassing lineages 6 and 9 (Marburger *et al.*, 2018). In chapter 2 we characterised and explored immune gene variation in a single toll like receptor (TLR2) across these nine lineages. Restriction site Associated DNA (RAD) sequence data previously identified that SNP ratios generally increased in higher lineages (with the exception of lineage 6 which was similar to lineage 9) (Marburger *et al.*, 2018). Similarly haplotype retentions in RAD data were lowest in lineage 1, highest in lineage 9 and variable in intermediate lineages (Marburger *et al.*, 2018). SNPs in TLR2 displayed a markedly different pattern to those observed in RAD data. SNPs were highest in lineage 7 and lineage 2 and lowest in lineages 1 and 6. In addition lineage 9 had four haplotypes at TLR2, lineage 1 and 6 had two haplotypes and data from the remaining lineages suggested a haplotype count between 6 and 10. When the possibility of SNP sharing between lineages was examined a cluster of SNPs shared between lineages 2 and 7 suggested potential haplotype sharing. This high frequency of SNP sharing between lineages that were relatively distantly related is unexpected and may be evidence of introgression, convergence or incomplete lineage sorting (Těšický and Vinkler, 2015). However, introgression was not supported by broader phylogenetic evidence from both immune gene and RAD data, while convergence, although theoretically possible, is considered unlikely because of the number of SNPs shared across multiple lineages. Incomplete lineage sorting was considered a more likely mechanism for this level of SNP sharing, especially if a force such as balancing selection was causing the persistence of shared SNPs across multiple lineages.

6.1.2 Variation of TLR1 and TLR2 across diploid and polyploid Corydoras populations

The Araguaia river *Corydoras* community is composed of three species, *C. maculifer* (diploid), *C. araguaiaensis* (putative tetraploid) and an un-described lineage 8 species, which will not be included further in this discussion. In chapter 3 we assessed the immune gene diversity of TLR1 and TLR2 across *C. maculifer* and *C. araguaiaensis* populations. In terms of SNP diversity *C. araguaiaensis*, the putative tetraploid, had significantly higher numbers of synonymous and non-synonymous SNPs in both TLR1 and TLR2, and individuals had a high likelihood of carrying unique SNP profiles even when only looking at the functionally significant non-synonymous SNPs. *Corydoras maculifer* (diploid) had a maximum of one SNP in either gene across all individuals. In addition, both short range SNP phasing and SNP read ratio scores indicated that *C. maculifer* had a maximum of two haplotypes for each TLR and *C. araguaiaensis* had as many as four haplotypes for each TLR.

Two interesting discussion points arise from these findings, the high diversity of *C. araguaiaensis* and the surprisingly low diversity of *C. maculifer*. The high diversity found in *C. araguaiaensis* supports the theory that polyploids are subjected to a higher degree of genetic drift and mutation retention due to additional haplotypes being effectively freed from selection. The high number of low frequency SNPs observed is supportive of this and the theory that selection may be diluted at higher ploidy levels should an advantageous mutation arise (Otto and Whitton, 2000). However, and somewhat conversely, these observations may also support the idea that mechanisms such as heterozygote advantage and negative frequency dependence occur at these immune genes (King, Seppälä and Neiman, 2012). However, the even spread of SNPs across all areas of the TLR genes does not confirm this, as under the heterozygote advantage and negative frequency dependence mechanisms one would expect advantageous diversity to be centred in the pathogen recognition receptor domain of the TLR (Medvedev, 2013). In *C. araguaiaensis* SNPs were relatively evenly spread at low frequency across both TLRs, a finding, which may favour the theory of weak selection and high levels of drift.

Intriguingly, the immune genes diversity of in *C. maculifer* was lower then expected even considering its diploid state. The numbers of SNPs recorded in *C. maculifer* in immune genes, which tend to be highly polymorphic (Netea, Wijmenga and O'Neill, 2012), are comparable to those recorded in threatened and/or bottleneck populations of other taxa e.g. birds (Grueber *et al.*, 2015; Gilroy *et al.*, 2017). It is possible that *C. maculifer's* recent evolutionary history also contains a bottleneck event, however we do not have the data necessary to confirm or refute this.

6.1.3 Parasite communities across diploid and polyploid Corydoras populations

Preliminary data (Childerstone & Taylor 2012, unpublished) indicated differences in parasite burden between the diploid *C. maculifer* and the putative tetraploid *C. araguaiaensis*. Chapter 4 explored these host parasite relationships further and examined potential links with the TLR immune gene diversity observed in Chapter 3. Parasite prevalence – the proportion of each species to carry an infection – was very similar between the two *Corydoras* host species. However, parasite intensity – the number of parasites found within infected individuals – was generally higher, although not significantly so, in *C. maculifer*. These combined findings suggested that members of both host species were equally likely to harbour some level of parasite infection but infection intensity was generally higher in *C. maculifer*. A trend was found between larger host body size and higher parasite burden. Because of this and because there were subtly different size demography's between the two host species, analyses were conducted looking at parasite abundance – the number of parasites per individual – while accounting for host size. With size accounted for parasite abundances were significantly higher in *C. maculifer,* suggesting that host species played a role in parasite abundance.

Analyses were performed to investigate whether overall TLR diversity had an impact on parasite abundance, or if specific SNPs were associated with parasite load. Both analyses found no relationships between SNP diversity or SNP parasite associations across TLR1 and TLR2. This may be an indication that there is genuinely no link between the immune gene diversity observed in *C. araguaiaensis* and its reduced parasite abundance, i.e. supporting the theory of multiple low frequency mutations arising due to diluted selection and increased drift. However it may also be that these TLRs do not play a major role in host parasite defence in *C. araguaiaensis*. Analysis of a greater range of immune genes would help to solve this question.

6.1.4 Characterising pathogen recognition receptors in the Corydoras maculifer genome

One of the findings from Chapter 3 indicated that TLR diversity was very low in *C. maculifer*, and was on a level with threatened species or populations, which had recently been through a bottleneck event (Grueber *et al.*, 2015; Gilroy *et al.*, 2017). Using genomic data for *C. maculifer* we attempted to isolate and characterise the remaining pathogen recognition receptors (PRRs) in a single individual to ascertain if variation was equally low across immune genes. In fish the PRR family is composed of the toll like receptors (TLRs), nucleotide-binding oligomerization domain (NOD) and leucine rich repeat containing receptors (NLRs) and the retinoic acid inducible gene 1 (RIG-1) like helicases (RHLs) (Aoki and Hirono, 2006). Of these three gene families five TLR genes (TLR1, TLR2, TLR7, TLR18 and TLR25) and eight NLRs (NOD1, NOD2 and six variants of NLRP3) were identified. Only one SNP was identified within the TLR family and

that was in TLR1 a SNP also documented in Chapter 3. Likewise NOD1 and NOD2 were equally devoid of variation, however the NLRP3 associated genes were considerably more variable but this variation was suspected to be a cumulative result of tandem duplication and genome assembly error. Simple Modular Architecture Research Tool (SMART) analysis found that the NLRP3 genes held fish specific NACHT associated (FISNA) domain, which is a structure commonly found in NLRs (Meng *et al.*, 2009; Rajendran *et al.*, 2012). Tandem duplications have previously been associated with NACHT regions in other species (Hamada *et al.*, 2012). Because of this and because *C. maculifer* is generally thought to be diploid, these additional haplotypes for the NLRP3 associated genes were accredited to potential tandem duplication events. Tandem duplication events can confound assembly processes and are frequently over or under represented across the assembled genome (Bailey *et al.*, 2004).

6.2 Further work

This thesis has explored the effects of whole genome duplication (WGD) events on immune gene diversity and parasite burden in the Corydoradinae catfishes. Little is known of the long-term impacts of WGD in animals and the Corydoradinae provide an excellent system for exploration, their mixed community structures allow for direct comparison between ploidy levels and their tendency to share habitats mean that environmental and pathogenic factors can be controlled for. This research has identified a number of potentially interesting observations. It has characterised TLR2 across the nine lineages, identifying the potential for incomplete lineage sorting between lineages 2 and 7. It has also found a much greater level of immune gene diversity and reduced parasite intensity in the putative tetraploid *C. araguaiaensis* when compared to the sympatrically coexisting diploid species *C. maculifer*. These results could be due to WGD derived high diversity and parasite resistance in *C. araguaiaensis,* or a bottlenecking event and high parasitic tolerance in *C. maculifer*, or a combination of both. The work here was unable to resolve between these possibilities. However it would appear that low immune gene diversity in *C. maculifer* is a trend that extends to other PRR families not just TLR1 and TLR2.

The research outlined in this thesis has been limited by a number of factors, which could be interesting points of exploration for future work. Firstly, the investigation of TLR2 across the nine lineages is based on single individual representations from seven of the nine lineages. Expansion of the sample size to include a wider range of individuals and species from these lineages, along with a greater variety of immune genes, would provide a better understanding and firmer grounding on the impacts of WGD on immune gene evolution and perhaps more broadly on impacts of WGD on evolution of animal systems.

Analysis across both Chapter 2 and Chapter 3 was also limited by our inability to fully phase individual haplotypes from the sequencing data. This is an intrinsic limitation associated with polyploid sequence data. There are a number of ways around these limitations however. One method would be to clone and Sanger sequence a range of different amplicons with primers designed to cover different sections of the immune genes in question. This would give a range of longer read lengths for individual haplotypes. Another method for achieving this aim would be to invest in longer ranged sequencing technologies such as PacBio or Nanopore MinION. Being able to identify full haplotype sequences would enable us to look for common haplotypes across species populations, as opposed to just looking at shared SNPs, and do a more effective selection analysis including deriving estimates of haplotype dN/dS ratios. It would also allow us to do a broader association test with the parasite data.

One of the main findings of Chapter 3 was that *C. araguaiaensis* appears to have retained four haplotype copies of TLR1 and TLR2. A stop codon was found in TLR1 in three individuals indicating that some level of gene silencing was occurring, however we do not know if copies are still functionally active in other individuals or genes. If these additional haplotypes are not transcribed then the observation of haplotype retention is inconsequential. Amplification of TLR genes from mRNA derived cDNA followed by sequencing would allow us to identify if all haplotypes are transcribed or if some are functionally redundant but have not been completely lost from the genome yet.

Parasite load was quantified across populations of *C. maculifer* and *C. araguaiaensis* in Chapter 4. Once again this analysis was limited by host and parasite sample size. In order to get a better resolution of parasite community's parasite sample sizes needed to be larger, which by extension means increasing host sample size. We also ran immune gene SNP association analyses in this chapter, which found no links between SNPs and parasite load. This may mean that there is genuinely no association between specific SNPs and parasite load in populations of *C. araguaiaensis*. Or other immune genes, that we have no data on, might be more closely associated with parasite burden. If the level of diversity observed in TLR1 and TLR2 extends to other immune genes it may be that these SNPs are more closely associated with the parasite abundances observed, in any case looking at a broader suite of immune genes would make an interesting comparison.

The final chapter of this thesis looked at the PRR immune gene family across the *C. maculifer* genome. The original aim of this chapter was to characterise PRRs across both *C. maculifer* and *C. araguaiaensis* genomes. However because the sequencing coverage of the *C. araguaiaensis* genome was poor no additional PRRs, beyond TLR1 and TLR2 could be isolated from the genomic data. Further sequencing of the *C. araguaiaensis* genome, ideally aiming at

longer sequence lengths (such as those offered by PacBio or Nanopore technologies) could be an effective solution to this problem and provide a further data source and species comparison.

Overall this thesis has identified a number of trends across *Corydoras* species regarding their immune gene diversity and parasite load. The Araguaia River community (composed of the two sympatric *Corydoras* species, *C. maculifer* and *C. araguaiaensis*) has provided a unique opportunity to compare the effects of whole genome duplication on immune gene diversity in a community that shares pathogenic and environmental exposures. This research has opened a number of questions in the fields of immune gene evolution and whole genome duplication, which will hopefully be further explored and resolved with future empirical studies.

References

Adams, K. L. and Wende, J. F. (2005) 'Novel patterns of gene expression in polyploid plants', *Trends in Genetics*, 21(10), pp. 539–543.

Alexandrou, M. a *et al.* (2011) 'Competition and phylogeny determine community structure in Müllerian co-mimics.', *Nature*. Nature Publishing Group, 469, pp. 84–89.

Alexandrou, M. and Taylor, M. (2011) *Evolution, ecology and taxonomy of the Corydoradinae revisited in Identifying Corydoradinae catfish (Supplement 1)*. Edited by I. A. . Fuller and H.-G. Evers. Cardiff: Ian Fuller Enterprises.

Aljanabi, S. M. and Martinez, I. (1997) 'Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques', *Nucleic Acids Research*, 25(22), pp. 4692–4693.

Alvarez-Pellitero, P. (2008) 'Fish immunity and parasite infections: from innate immunity to immunoprophylactic prospects.', *Veterinary immunology and immunopathology*, 126, pp. 171–198.

Aoki, T. and Hirono, I. (2006) 'Immune relevant genes of Japanese flounder, Paralichthys olivaceus.', *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, 1, pp. 115–121.

Aulchenko, Y. S. *et al.* (2007) 'GenABEL: An R library for genome-wide association analysis', *Bioinformatics*, 23(10), pp. 1294–1296.

Bailey, J. A. *et al.* (2004) 'Analysis of segmental duplications and genome assembly in the mouse', *Genome Research*, 14(5), pp. 789–801.

Benfey, T. J. (1999) 'The physiology and behavior of triploid fishes', *Reviews in Fisheries Science*, 7, pp. 39–67.

Berthelot, C. *et al.* (2014) 'The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates', *Nature communications*, 5, pp. 1–10.

Bonaparte, C. (1838) 'Synopsis vertebratorum systematis.', *Nuovi Annali delle Scienze Naturali (Bologna)*, 2, pp. 105–133.

Button, K. S. *et al.* (2013) 'Power failure: Why small sample size undermines the reliability of neuroscience', *Nature Reviews Neuroscience*. Nature Publishing Group, 14(5), pp. 365–376.

Carius, H. ., Little, T. . and Ebert, D. (2001) 'Genetic variation in a host-parasite association: potential for coevolution and frequency-dependent selection', *Evolution*, 55, pp. 1136–1145.

Cavalier-Smith, T. (1978) 'Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth tate, and the solution of the DNA C-value paradox', *Journal of Cell Science*, 34, pp. 247–278.

Chang, M. *et al.* (2011) 'Expression and functional characterization of the RIG-I-like receptors MDA5 and LGP2 in Rainbow trout (Oncorhynchus mykiss).', *Journal of virology*, 85, pp. 8403–8412.

Chen, S. N., Zou, P. F. and Nie, P. (2017) 'Retinoic acid-inducible gene I (RIG-I)-like receptors (RLRs) in fish: current knowledge and future perspectives', *Immunology*, 151(1), pp. 16–25.

Chen, Z. J. (2007) 'Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids', *Annual Review of Plant Biology*, 58(1), pp. 377–406.

Childerstone, A. and Taylor, M. (2012) *Determining fitness of the polyploidy catfish Corydoras araguaiaensis by comparing it to a diploid mimic Corydoras maculifer: using parsite burden as an indicator of fitness*. Bangor University.

Ching, B. *et al.* (2010) 'Transcriptional differences between triploid and diploid Chinook salmon (Oncorhynchus tshawytscha) during live Vibrio anguillarum challenge.', *Heredity*. Nature Publishing Group, 104, pp. 224–234.

Clavijo, B. J. *et al.* (2017) 'W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data', *bioRxiv*, pp. 1–12.

Coate, J. E. and Doyle, J. J. (2010) 'Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid.', *Genome biology and evolution*, 2, pp. 534–46.

Comai, L. (2005) 'The advantages and disadvantages of being polyploid', *Nature Rev. Genet.* Nature Publishing Group, 6, pp. 836–846.

Cronn, R. C., Small, R. L. and Wendel, J. F. (1999) 'Duplicated genes evolve independently after polyploid formation in cotton', *Proceedings of the National Academy of Sciences*, 96(25), pp. 14406–14411.

Dehal, P. and Boore, J. L. (2005) 'Two rounds of whole genome duplication in the ancestral vertebrate.', *PLoS biology*, 3, pp. 1700–1708.

Doherty, P. and Zinkernagel, R. (1975) 'Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex', *Nature*, 256, pp. 50–52.

Duchemin, M. B., Fournier, M. and Auffret, M. (2007) 'Seasonal variations of immune parameters in diploid and triploid Pacific oysters, Crassostrea gigas (Thunberg)', *Aquaculture*, 264, pp. 73–81.

Evans, D. *et al.* (1984) 'Nonspecific cytotoxic cells in fish (Ictalurus punctatus)', *Developmental*

and Comparative Immunology, 8, pp. 823–833.

Evans, D. L. *et al.* (1984) 'Nonspecific Cytotoxic Cells in Fish (Ictalurus Punctatus)', *Developmental & Comparative Immunology*, 8, pp. 303–312.

Fink, I. R. *et al.* (2016) 'Molecular and functional characterization of Toll-like receptor (Tlr)1 and Tlr2 in common carp (Cyprinus carpio)', *Fish and Shellfish Immunology*. Elsevier Ltd, 56, pp. 70–83.

Float, K. and Whitham, T. (1993) 'The "Hybrid Bridge" hypothesis: host shifting via plant hybrid swarms', *The American Naturalist*, 141, pp. 651–662.

Frankham, R., Ballou, J. and Briscoe, D. (2010) *Introduction to Conservation Genetics*. Second edi. Cambridge: Cambridge University Press.

Fritz, J. *et al.* (2007) 'Innate immune recognition at the epithelial barrier drives adaptive immunity: APCs take the back seat', *Trends in Immunology*, 29, pp. 41–49.

Fuller, I. A. and Evers, H.G. (2005) *Identifying Corydoradinae catfish*. Cardiff: Ian Fuller Enterprises.

Furlong, R. and Holland, P. W. (2004) 'Polyploidy and vertebrate ancestry: Ohno and beyond', *Biological Journal of the Linnean Society*, 82, pp. 425–430.

Garrison, E. and Marth, G. (2012) 'Haplotyp-based variant detection from short-read sequencing', *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

Gilroy, D. L. *et al.* (2017) 'Toll-like receptor variation in the bottlenecked population of the endangered Seychelles warbler', *Animal Conservation*, 20(3), pp. 235–250.

Gorelick, R. and Olson, K. (2013) 'Polyploidy is genetic hence may cause non-adaptive radiations, whereas Pseudopolyploidy is genomic hence may cause adaptive non-radiations', *Journal of Experimental Zoology: Molecular and Developmental Evolution*, 320, pp. 286–294.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome.', *Nature biotechnology*, 29(7), pp. 644–652.

Grueber, C. E. *et al.* (2015) 'Toll-like receptor diversity in 10 threatened bird species: relationship with microsatellite heterozygosity', *Conservation Genetics*. Springer Netherlands, 16(3), pp. 595–611.

Hamada, M. *et al.* (2012) 'The complex NOD-like receptor repertoire of the coral acropora digitifera includes novel domain combinations', *Molecular Biology and Evolution*, 30(1), pp. 167–176.

Hill, A. V. S. (1999) 'Defence by diversity', *Nature*, 398, pp. 668–669.

Hoang, D. T. *et al.* (2018) 'UFBoot2: Improving the ultrafast bootstrap approximation', *Molecular Biology and Evolution*, 35(2), pp. 518–522.

Hughes, A. L. and Piontkivska, H. (2008) 'Functional diversification of Toll-like Receptor gene family', *Immunogenetics*, 60(5), pp. 249–256.

Humann, J . *et al.* (2018) *GenSAS c5.1: A Web-Based Platform for Structural and Functional Annotation and Curation of Genomes.*, *Next Generation Genome Annotation and Analysis Workshop, International Pland & Animal Genome Conference XXVI, San Diego, CA, USA*.

Jackson, J. a. and Tinsley, R. C. (2003) 'Parasite infectivity to hybridising host species: A link between hybrid resistance and allopolyploid speciation?', *International Journal for Parasitology*, 33, pp. 137–144.

Jones, S. R. (2001) 'The occurrence and mechanisms of innate immunity against parasites in fish', *Developmental and comparative immunology*, 25, pp. 841–852.

Keller, M. J. and Gerhardt, H. C. (2001) 'Polyploidy alters advertisement call structure in gray treefrogs.', *Proceedings. Biological sciences / The Royal Society*, 268, pp. 341–5.

Khan, R. A. (2012) 'Host-parasite interactions in some fish species', *Journal of Parasitology Research*, 2012, pp. 1–7.

Kihara, H. and Ono, T. (1926) 'Chromosomenzahlen und Systematische Grruppierung der Rumex-Arten', *Zeitschrift für Zellforschung und mikroskopische Anatomie*, 4, pp. 475–481.

King, K. C., Seppälä, O. and Neiman, M. (2012) 'Is more better? Polyploidy and parasite resistance.', *Biology letters*, 8, pp. 598–600.

Koskella, B. and Lively, C. M. (2009) 'Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode.', *Evolution; international journal of organic evolution*, 63, pp. 2213–21. doi: 10.1111/j.1558-5646.2009.00711.x.

Kuzmin, Y. *et al.* (2011) 'Camallanus Railliet et Henry, 1915 (Nematoda, Camallanidae) from Australian freshwater turtles with descriptions of two new species and molecular differentiation of known taxa', *Acta Parasitologica*, 56(2), pp. 213–226.

Lafferty, K. D., Dobson, A. P. and Kuris, A. M. (2006) 'Parasites dominate food web links.', *Proceedings of the National Academy of Sciences*, 103, pp. 11211–11216.

Langston, A., Johnstone, R. and Ellis, A. (2001) 'The kinetics of the hypoferraemic response and changes in levels of alternative complement activity in diploid and triploid Atlantic salmon, following injection of lipopolysaccharide.', *Fish & shellfish immunology*, 11, pp. 333–45.

Lester, R. J. G. and McVinish, R. (2016) 'Does moving up a food chain increase aggregation in parasites?', *Journal of the Royal Society Interface*, 13(118), pp. 1–11.

Letunic, I. and Bork, P. (2018) '20 years of the SMART protein domain annotation resource', *Nucleic Acids Research*. Oxford University Press, 46, pp. D493–D496.

Li, H. and Durbin, R. (2009) 'No TitleFast and Accurate short read alignment with Burrows-Wheeler Transform', *Bioinformatics*, 25, pp. 1754–1760.

Liu, Z. *et al.* (2016) 'The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts', *Nature Communications*. Nature Publishing Group, 7, pp. 1–13.

Lively, C. M. *et al.* (2004) 'Host sex and local adaptation by parasites in a snail-trematode interaction.', *The American naturalist*, 164, pp. S6–S18.

Lo, C. M., Morand, S. and Galzin, R. (1998) 'Parasite diversity\host age and size relationship in three coral-reef fishes from French Polynesia', *International Journal for Parasitology*, 28(11), pp. 1695–1708.

Love, A. and Love, D. (1943) 'The significance of differences in the distribution of diploids and polyploids', *Hereditas*, 29, pp. 145–163.

Mable, B. K. (2004) "Why polyploidy is rarer in animals then in plants": myths and mechanisms', *Biological Journal of the Linnean Society*, 82, pp. 453–466.

Mable, B. K., Alexandrou, M. a. and Taylor, M. I. (2011) 'Genome duplication in amphibians and fish: an extended synthesis', *Journal of Zoology*, 284, pp. 151–182.

Magnadóttir, B. (2006) 'Innate immunity of fish (overview).', *Fish & shellfish immunology*, 20, pp. 137–151.

Marburger, S. (2015) *Investigating Mechanisms of Genome Expansion in Corydoradinae catfish*. University of Bangor.

Marburger, S. *et al.* (2018) 'Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes', *Proceedings of the Royal Society B*, 285, pp. 1–10.

Masterson, J. (1994) 'Stomatal size in fossil plants : evidence for polyploidy in majority of angiosperms', *Science*, 264, pp. 421–424.

Medvedev, A. E. (2013) 'Toll-Like Receptor Polymorphisms, Inflammatory and Infectious Diseases, Allergies, and Cancer', *Journal of Interferon & Cytokine Research*, 33(9), pp. 467–484.

Medzhitov, R. and Janeway, C. (2002) 'Decoding the Patterns of Self and Nonself by the Innate Immune System', *Science*, 296, pp. 298–300.

Meng, G. *et al.* (2009) 'A Mutation in the Nlrp3 Gene Causing Inflammasome Hyperactivation Potentiates Th17 Cell-Dominant Immune Responses', *Immunity*. Elsevier Ltd, 30(6), pp. 860–874.

Netea, M. G., Wijmenga, C. and O'Neill, L.A.J. (2012) 'Genetic variation in Toll-like receptors and disease susceptibility.', *Nature immunology*, 13(6), pp. 535–42.

Neumann, N. *et al.* (2001) 'Antimicrobial mechanisms of fish phagocytes and their role in host defense', *Developmental and Comparative Immunology*, 25, pp. 807–825.

Nguyen, L. T. *et al.* (2015) 'IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution*, 32(1), pp. 268–274.

Nijssen, H. (1970) *Revision of the Surinam catfishes of the genus Corydoras Lacepede, 1803 (Pices, Siluriformes, Callichthyidae)*. University of Amsterdam.

Nijveen, H. *et al.* (2013) 'QualitySNPng: a user-friendly SNP detection and visualization tool.', *Nucleic acids research*, 41(Web Server issue), pp. 587–590.

Noreen, M. and Arshad, M. (2015) 'Association of TLR1, TLR2, TLR4, TLR6 and TIRAP polymorphisms with disease susceptibility', *Immunologic Research*. Springer US, 62, pp. 234–252.

Nuismer, S. L. and Otto, S. P. (2004) 'Host – parasite interactions and the evolution of ploidy', *Proceedings of the National Academy of Sciences*, 101, pp. 11036–11039.

Oliveira, C. *et al.* (1992) 'Extensive chromosomal rearrangements and nuclear DNA content changes in the evolution of the armoured catfishes genus Corydoras (Pisces, Siluriformes, Callichthyidae)', *Journal of Fish Biology*, 40, pp. 419–431.

Orr, H. A. (1990) '"Why polyploidy is rarer in animals than in plants" revisited', *The American Naturalist*, 136, pp. 759–770.

Osnas, E. E. and Lively, C. M. (2006) 'Host ploidy, parasitism and immune defence in a coevolutionary snail-trematode system.', *Journal of evolutionary biology*, 19, pp. 42–48.

Otto, S. P. and Goldstein, D. . (1992) 'Recombination and the evolution of diploidy', *Genetics*, 131, pp. 745–751.

Otto, S. P. and Whitton, J. (2000) 'Polyploid incidence and evolution', *Annual Review of Genetics*, 34, pp. 401–437.

Paquin, C. and Adams, J. (1983) 'Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations', *Nature*, 302, pp. 495–500.

Pasquier, J. *et al.* (2016) 'Gene evolution and gene expression after whole genome duplication in fish : the PhyloFish database', *BMC Genomics*. BMC Genomics, 17(368), pp. 1–10.

Peer, Y. Van De, Maere, S. and Meyer, A. (2009) 'The evolutionary significance of ancient genome duplications', *Nature Reviews, Genetics*, 10, pp. 725–732.

Pereira e Silva, J., Furtado, A. P. and Nascimento dos Santos, J. (2014) 'Proteomic profile of Ortleppascaris sp.: A helminth parasite of Rhinella marina in the Amazonian region', *International Journal for Parasitology: Parasites and Wildlife*. Australian Society for

Parasitology, 3(2), pp. 67–74.

Peterson, B. K. *et al.* (2012) 'Double Digest RADseq : An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species', *PloS one*, 7(5), pp. 1–11.

Petit, J. *et al.* (2004) 'Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype', *Nature*, 431, pp. 946–957.

Phillips, K. P. *et al.* (2018) 'Immunogenetic novelty confers a selective advantage in host–pathogen coevolution', *Proceedings of the National Academy of Sciences*, 10, pp. 1–6.

Pietretti, D. and Wiegertjes, G. F. (2014) 'Ligand specificities of Toll-like receptors in fish: Indications from infection studies', *Developmental and Comparative Immunology*. Elsevier Ltd, 43(2), pp. 205–222.

Poulin, R. (2000) 'Variation in the intraspecific relationship between fish length and intensity of parasitic infection: Biological and statistical causes', *Journal of Fish Biology*, 56(1), pp. 123–137.

Poulin, R. (2007) *Evolutionary Ecology of Parasites*. Princeton: Princeton University Press.

Prosser, S. W. J. *et al.* (2013) 'Advancing nematode barcoding: A primer cocktail for the cytochrome c oxidase subunit I gene from vertebrate parasitic nematodes', *Molecular Ecology Resources*, 13, pp. 1108–1115.

Råberg, L., Graham, A. L. and Read, A. F. (2009) 'Decomposing health: Tolerance and resistance to parasites in animals', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513), pp. 37–49.

Rajendran, K. V. *et al.* (2012) 'Pathogen recognition receptors in channel catfish: I. Identification, phylogeny and expression of NOD-like receptors', *Developmental and Comparative Immunology*. Elsevier Ltd, 37, pp. 77–86.

Ramsey, J. and Schemske, D. W. (1998) 'Pathways, mechanisms, and rates of polyploid formation in flowering plants', *Annual Review of Ecology, Evolution and Systematics*. Nature Publishing Group, 29, pp. 467–501.

Reite, O. B. (1998) 'Mast cells / eosinophilic granule cells of teleostean fish : a review focusing on staining properties and functional responses', *Fish & Shellfish Immunology*, 8, pp. 489–513.

Revell, L. J. (2012) 'phytools: An R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution*, 3(2), pp. 217–223.

Robertson, F. M. *et al.* (2017) 'Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification', *Genome Biology*. Genome Biology, 18(1), pp. 1–14.

Rocha, J. and Chiarini-Garcia, H. (2007) 'Mast cell heterogeneity between two different species

of Hoplias sp. (Characiformes: Erythrinidae): Response to fixatives, anatomical distribution, histochemical contents and ultrastructural features', *Fish and Shellfish Immunology*, 22, pp. 218–229.

Ryce, E. K. N., Zale, A. V and MacConnell, E. (2004) 'Effects of fish age and development of whirling parasite dose on the disease in rainbow trout', *Diseases of Aquatic Organisms*, 59, pp. 225–233.

Salaun, B., Romero, P. and Lebecque, S. (2007) 'Toll-like receptor's two-edged sword: When immunity meets apoptosis', *European Journal of Immunology*, 37, pp. 3311–3318.

Santini, F. *et al.* (2009) 'Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes.', *BioMedCentral Evolutionary Biology*, 9, pp. 1–15.

Sapp, S. G. H. *et al.* (2017) 'Beyond the raccoon roundworm: The natural history of non-raccoon Baylisascaris species in the New World', *International Journal for Parasitology: Parasites and Wildlife*. Elsevier Ltd, 6(2), pp. 85–99.

Scapigliati, G. (2013) 'Functional aspects of fish lymphocytes', *Developmental and Comparative Immunology*. Elsevier Ltd, 41, pp. 200–208.

Simao, F. A. *et al.* (2015) 'Genome analysis BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212.

Šimková, A. *et al.* (2013) 'MHC genes and parasitism in Carassius gibelio, a diploid-triploid fish species with dual reproduction strategies.', *BioMedCentral Evolutionary Biology*, 13, pp. 2–15.

Skevaki, C. *et al.* (2015) 'Single nucleotide polymorphisms of Toll-like receptors and susceptibility to infectious diseases', *Clinical and Experimental Immunology*, 180(2), pp. 165–177.

Slade, R. W. and McCallum, H. I. (1992) 'Overdominant Vs. frequency-dependent selection at MHC loci', *Genetics Society of America*, 132, pp. 861–864.

Solbakken, M. H. *et al.* (2017) 'Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system', *Proceedings of the Royal Society B: Biological Sciences*, 284, pp. 1–9.

Solbakken, M. H. *et al.* (2018) 'Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system', *Proceedings of the Royal Society B: Biological Sciences*, 284(1853), pp. 1–9.

Soltis, D. ., Visger, C. . and Soltis, P. . (2014) 'The polyploidy revolution... and now: Stebbins revisited', *American Journal of Botany*, 101, pp. 1057–1078.

Soltis, P. S. *et al.* (2015) 'Polyploidy and genome evolution in plants', *Current Opinion in Genetics and Development*. Elsevier Ltd, 35, pp. 119–125.

Song, K. *et al.* (1995) 'Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(17), pp. 7719–7723.

Spurgin, L. G. *et al.* (2011) 'Gene conversion rapidly generates major histocompatibility complex diversity in recently founded bird populations', *Molecular Ecology*, 20, pp. 5213–5225.

Spurgin, L. G. and Richardson, D. S. (2010) 'How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings.', *Proceedings. Biological sciences / The Royal Society*, 277, pp. 979–88.

Stebbins, G. L. (1940) 'The significance of polyploidy in plant evolution', *The American Naturalist*, 74, pp. 54–66.

Stone, K. ., Prussin, C. and Metcalfe, D. . (2010) 'IgE, mast cells, basophils, and eosinophils', *Journal of Allergy and Clinical Immunology*. Elsevier Ltd, 125, pp. S73–S80.

Strube, C., Heuer, L. and Janecek, E. (2013) 'Toxocara spp. infections in paratenic hosts', *Veterinary Parasitology*. Elsevier B.V., 193(4), pp. 375–389.

Sunnucks, P. and Hales, D. F. (1996) 'Numerous transposed sequences of mitochondrial cytochrome oxidase 1-11 in aphids of the genus Sitobion (Hemiptera: Aphididae)', *Molecular Biology and Evolution*, 13(August), pp. 510–524.

Szostakowska, B., Myjak, P. and Kur, J. (2002) 'Identification of anisakid nematodes from the Southern Baltic Sea using PCR-based methods', *Molecular and Cellular Probes*, 16(2), pp. 111–118.

Takeda, K. and Akira, S. (2005) 'Toll-like receptors in innate immunity', *International Immunology*, 17(1), pp. 1–14.

Temperley, N. D. *et al.* (2008) 'Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss.', *BMC genomics*, 9, p. 62.

Těšický, M. and Vinkler, M. (2015) 'Trans-Species Polymorphism in Immune Genes: General Pattern or MHC-Restricted Phenomenon?', *Journal of Immunology Research*, 2015, pp. 1–10.

Thompson, J. and Lumaret, R. (1992) 'The evolutionary dynamics of polyploid plants: origins, establishment and persistance', *Trends in Ecology & Evolution*, 7, pp. 302–307.

Tiwary, B. K., Kirubagaran, R. and Ray, A. K. (2005) 'The biology of triploid fish', *Reviews in Fish Biology and Fisheries*, 14, pp. 391–402.

Uribe, C. *et al.* (2011) 'Innate and adaptive immunity in teleost fish : a review', *Veterinarni*

Medicina, 56, pp. 486–503.

Vale, M. (2008) 'The physiology of triploid fish: current knowledge and comparisons with diploid fish', *Fish and Fisheries*, 9, pp. 67–78.

Wagner, A. (2005) 'Energy constraints on the evolution of gene expression', *Molecular Biology and Evolution*, 22(6), pp. 1365–1374.

Waterhouse, R. M. *et al.* (2018) 'BUSCO applications from quality assessments to gene prediction and phylogenomics', *Molecular Biology and Evolution*, 35(3), pp. 543–548.

Weiss, R., Kukora, J. . and Adams, J. (1975) 'The relationship between enzyme activity, cell geometry and fitness in Saccharomyces cerevisiae', *Proceedings of the National Academy of Sciences*, 72, pp. 794–798.

Whyte, S. K. (2007) 'The innate immune response of finfish-a review of current knowledge.', *Fish & shellfish immunology*, 23(6), pp. 1127–1151.

Yandell, M. and Ence, D. (2012) 'A beginner ' s guide to eukaryotic genome annotation', *Nature Reviews, Genetics*. Nature Publishing Group, 13, pp. 329–342.

Zapata, A. *et al.* (2006) 'Ontogeny of the immune system of fish.', *Fish & shellfish immunology*, 20, pp. 126–136. doi: 10.1016/j.fsi.2004.09.005.

Zhao, F. *et al.* (2013) 'Expression profiles of toll-like receptors in channel catfish (Ictalurus punctatus) after infection with Ichthyophthirius multifiliis.', *Fish & shellfish immunology*, 35, pp. 993–997.

Zhou, L. and Gui, J. (2017) 'Natural and artificial polyploids in aquaculture', *Aquaculture and Fisheries*, 2(3), pp. 103–111.

Zinkernagal, R. . and Doherty, P. C. (1974) 'Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngenic or semiallogenic system', *Nature*, 248, pp. 701–702.

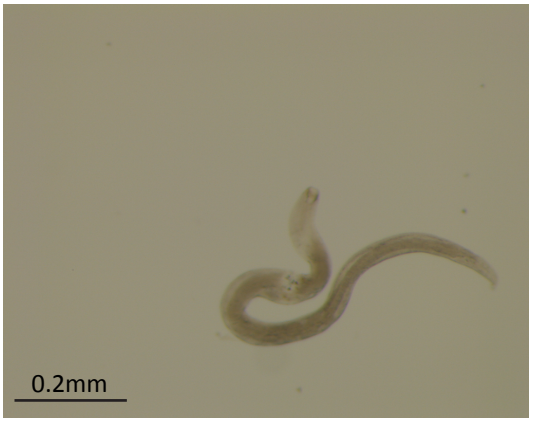
Appendix

Sample: 0005_1, Host: 2015_0005
ID: acanthocephalon, Magnification : 8.0x

Sample: 0005_9, Host: 2015_0005
ID: acanthocephalon, Magnification : 8.0x

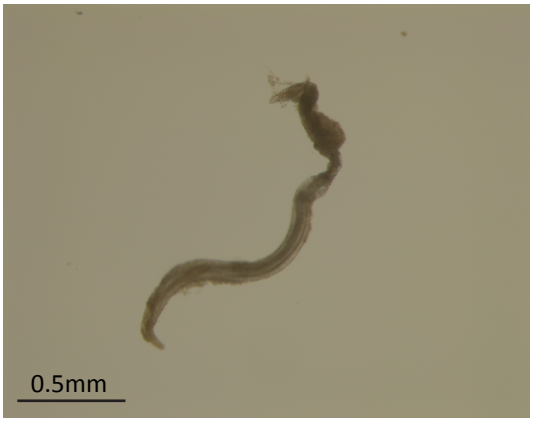
Sample: 0005_7, Host: 2015_0005
ID: nematode, Magnification : 8.0x

Sample: 0005_8, Host: 2015_0005
ID: nematode, Magnification : 5.0x

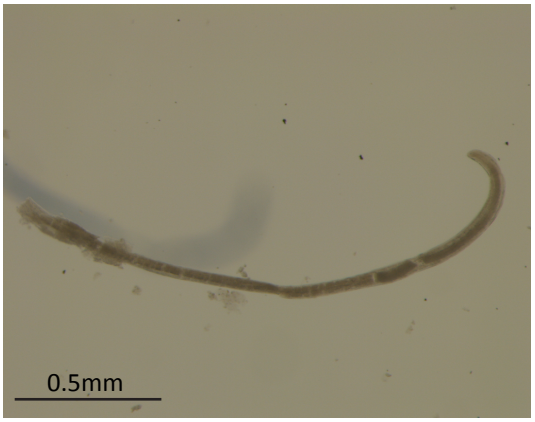


Sample: 0006_2, Host: 2015_0006
ID: nematode, Magnification : 12.0x

Sample: 0008_4, Host: 2015_0008
ID: nematode, Magnification : 5.0x

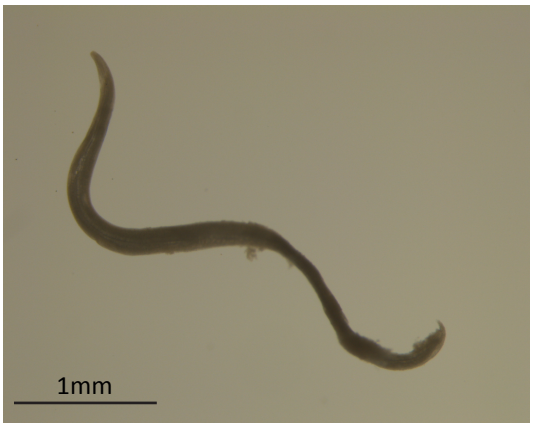


Sample: 0006_2, Host: 2015_0006
ID: nematode, Magnification : 5.0x

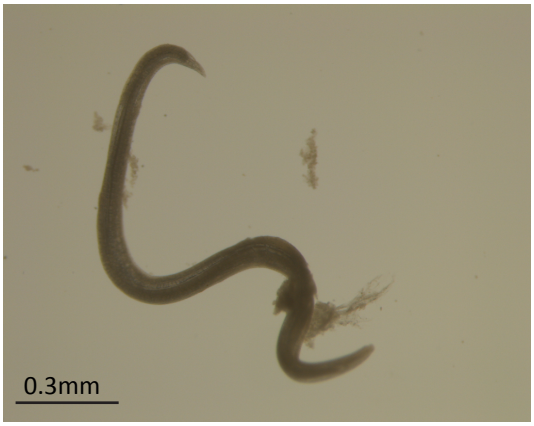


Sample: 0016_7, Host: 2015_0016
ID: nematode, Magnification : 6.3x

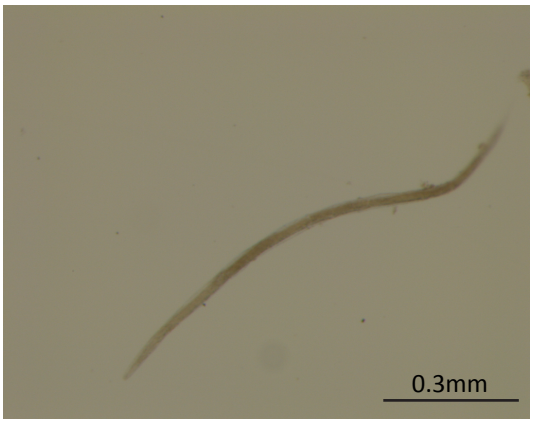
Sample: 0020_4, Host: 2015_0020
ID: nematode, Magnification : 12.0x



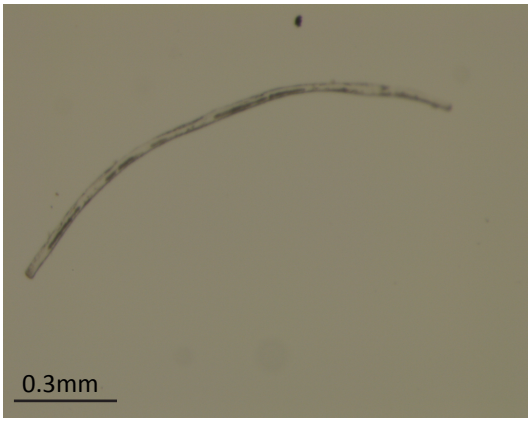
Sample: 0020_5, Host: 2015_0020
ID: nematode, Magnification : 3.2x



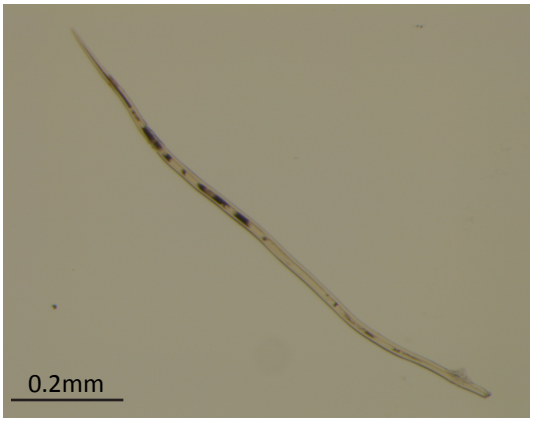
Sample: 0025_3, Host: 2015_0025
ID: nematode, Magnification : 8.0x



Sample: 0026_5, Host: 2015_0026
ID: nematode, Magnification : 10.0x



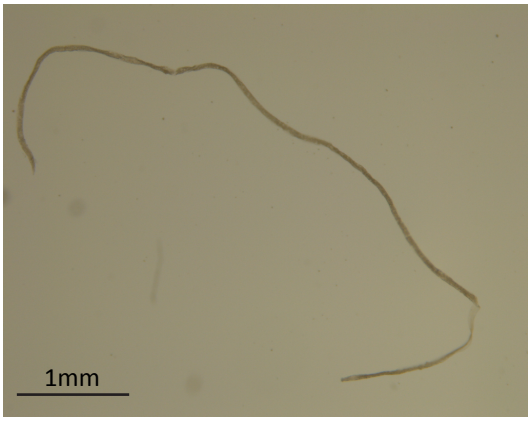
Sample: 0027_3, Host: 2015_0027
ID: nematode, Magnification : 8.0x



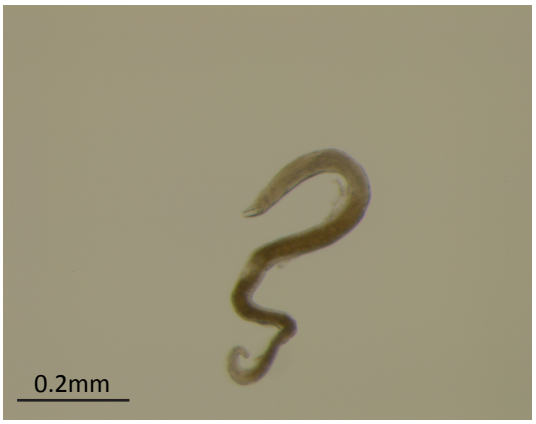
Sample: 0027_5, Host: 2015_0027
ID: nematode, Magnification : 12.0x

Sample: 0028_5, Host: 2015_0028
ID: nematode, Magnification : 5.0x

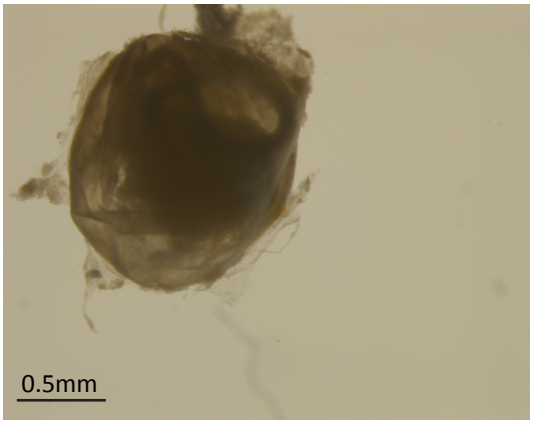
Sample: 0030_2, Host: 2015_0030
ID: nematode, Magnification : 5.0x



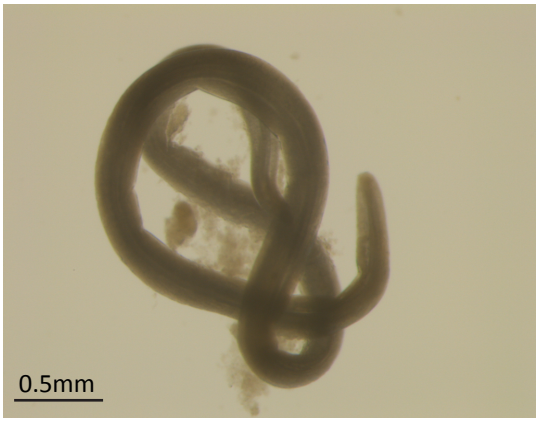
Sample: 0033_1, Host: 2015_0033
ID: nematode, Magnification : 2.5x



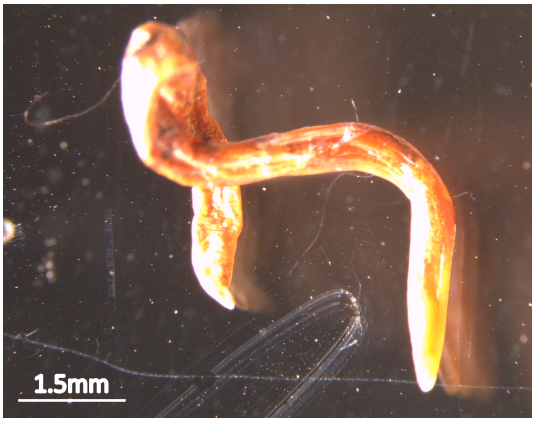
Sample: 0033_2, Host: 2015_0033
ID: nematode, Magnification : 12.0x



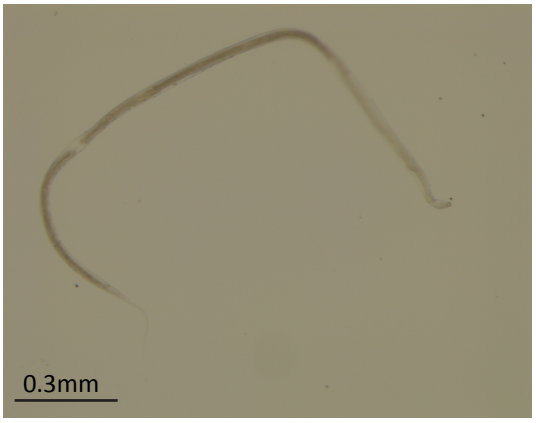
Sample: 0036_8, Host: 2015_0036
ID: nematode, Magnification : 4.0x



Sample: 0036_8, Host: 2015_0036
ID: nematode, Magnification : 4.0x

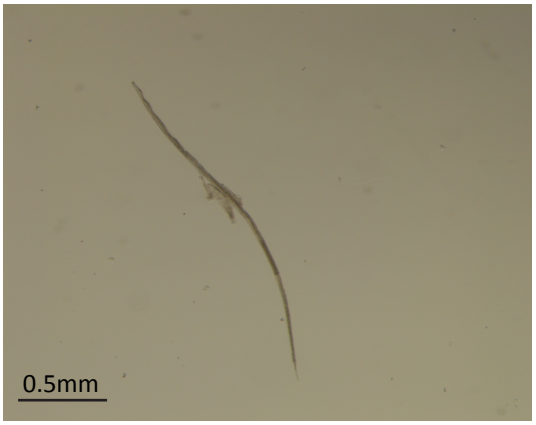


Sample: 0101_4, Host: 2015_0101
ID: nematode, Magnification : 1.6x

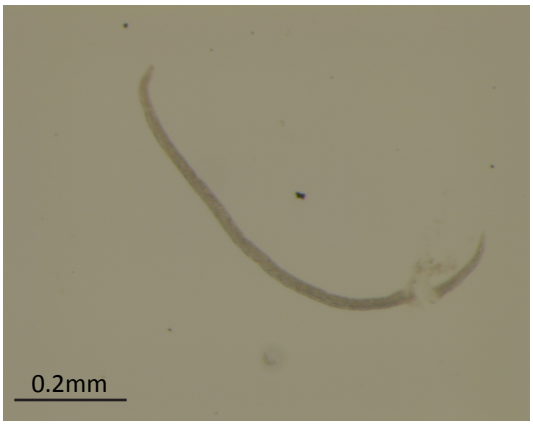


Sample: 304_4, Host: 2012_304
ID: nematode, Magnification : 8.0x

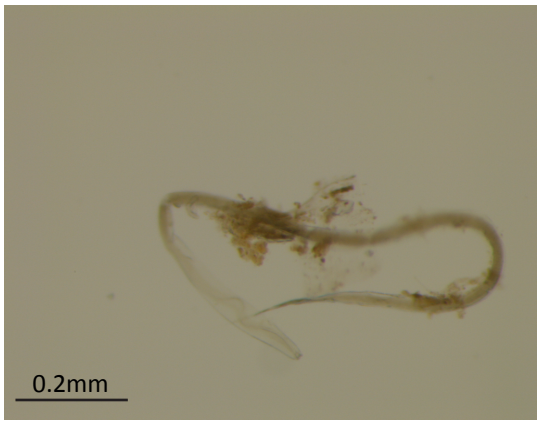
Sample: 305_2, Host: 2012_305
ID: nematode, Magnification : 5.0x



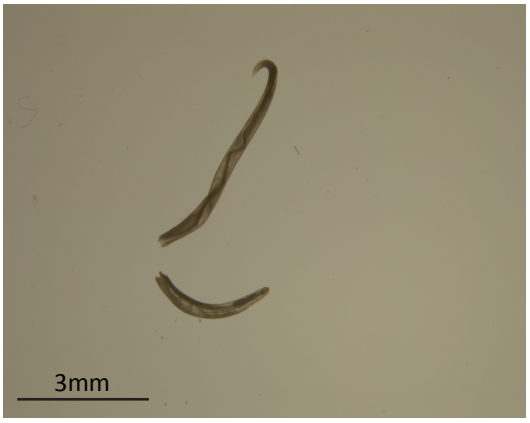
Sample: 306_4, Host: 2012_306
ID: nematode, Magnification : 4.0x



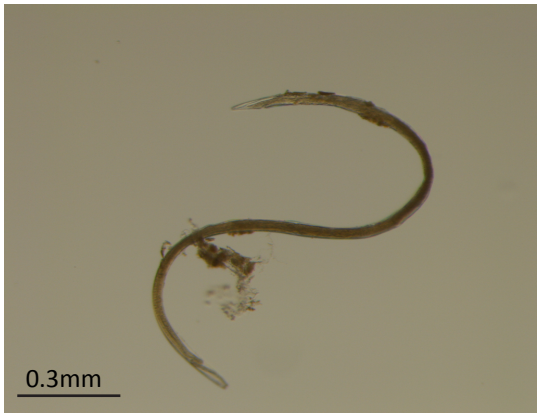
Sample: 306_5, Host: 2012_306
ID: nematode, Magnification : 12.0x



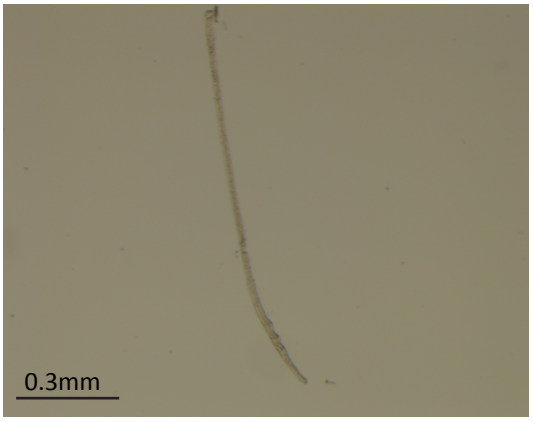
Sample: 308_3, Host: 2012_308
ID: nematode, Magnification : 12.0x



Sample: 309_2, Host: 2012_309
ID: nematode, Magnification : 1.0x

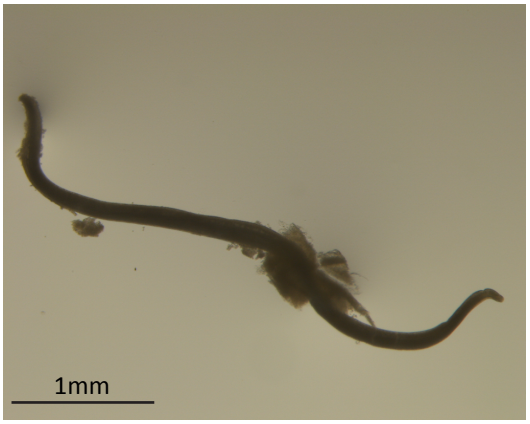


Sample: 311_2, Host: 2012_311
ID: nematode, Magnification : 8.0x

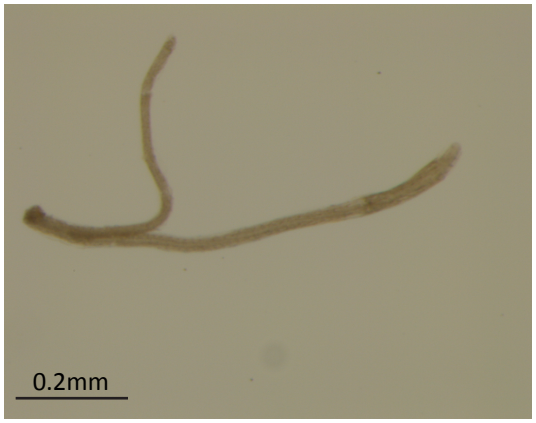


Sample: 311_3, Host: 2012_311
ID: nematode, Magnification : 8.0x

Sample: 312_4, Host: 2012_312
ID: nematode, Magnification : 12.0x

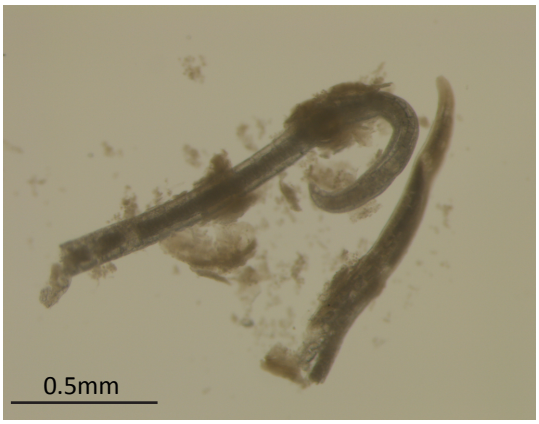


Sample: 315_2, Host: 2012_315
ID: nematode, Magnification : 3.2x

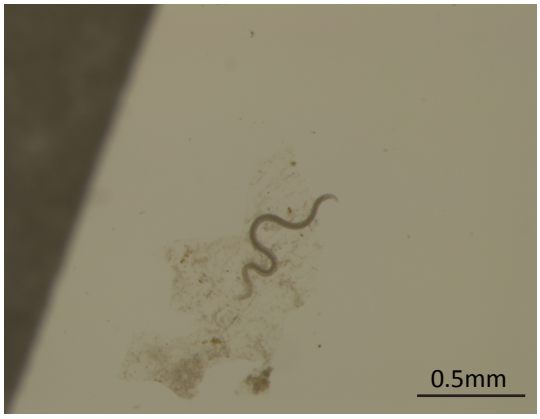


Sample: 316_5, Host: 2012_316
ID: nematode, Magnification : 12.0x

Sample: 323_5, Host: 2012_323
ID: nematode, Magnification : 8.0x

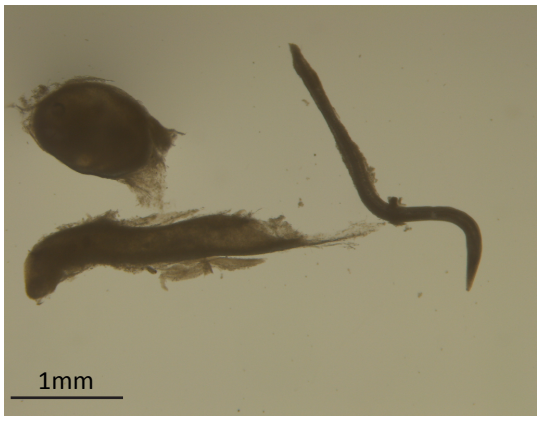


Sample: 323_6, Host: 2012_323
ID: nematode, Magnification : 6.3x



Sample: 323_8, Host: 2012_323
ID: nematode, Magnification : 5.0x

Sample: 324_4, Host: 2012_324
ID: nematode, Magnification : 5.0x

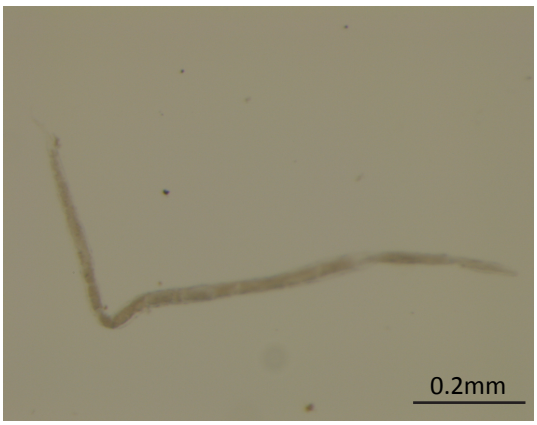


Sample: 325_3, Host: 2012_325
ID: nematode, Magnification : 2.5x

Sample: 325_5, Host: 2012_325
ID: nematode, Magnification : 5.0x

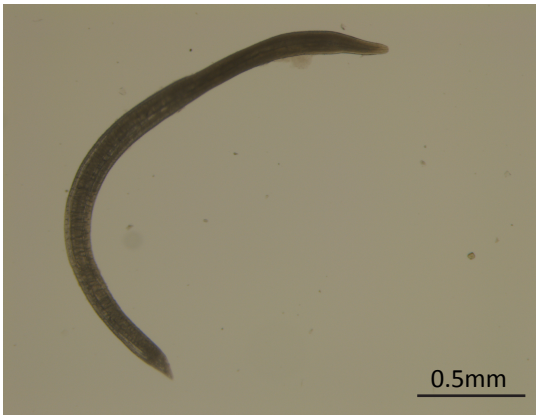
Sample: 325_6, Host: 2012_325
ID: nematode, Magnification : 5.0x

Sample: 327_8, Host: 2012_327
ID: nematode, Magnification : 5.0x



Sample: 327_11, Host: 2012_327
ID: nematode, Magnification : 12.0x

Sample: 333_2, Host: 2012_333
ID: nematode, Magnification : 5.0x

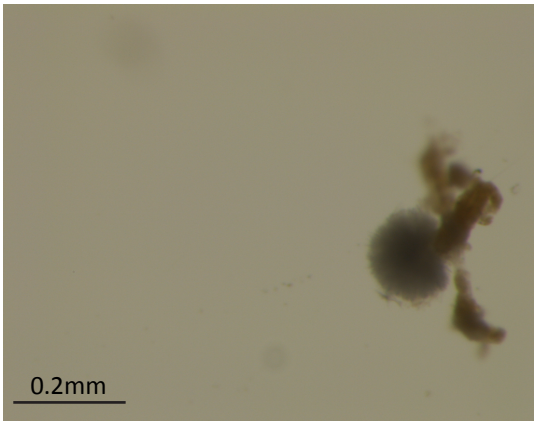


Sample: 334_4, Host: 2012_334
ID: nematode, Magnification : 5.0x

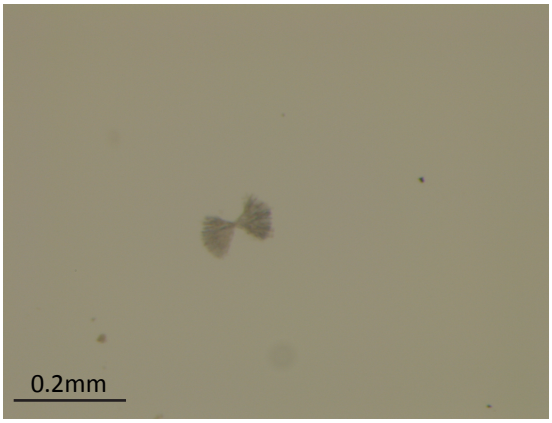
Sample: 339_5, Host: 2012_339
ID: nematode, Magnification : 9.0x

Sample: 0018_3, Host: 2015_0018
ID: unknown, Magnification : 12.0x

Sample: 0020_1, Host: 2015_0020
ID: isopode, Magnification : 2.5x



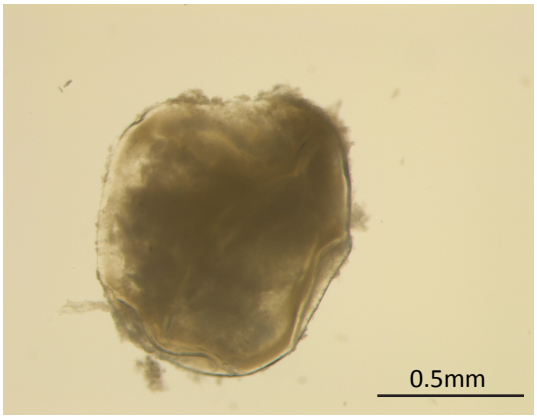
Sample: 0034_2, Host: 2015_0034
ID: possible fungi, Magnification : 12.0x



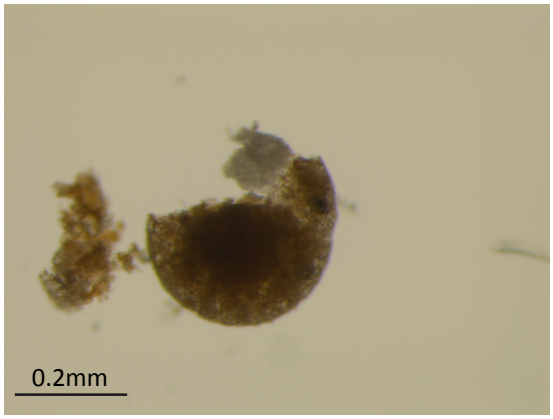
Sample: 0034_2, Host: 2015_0034
ID: possible fungi, Magnification : 12.0x

Sample: 0036_2, Host: 2015_0036
ID: cyst, Magnification : 12.0x

Sample: 0042_2, Host: 2015_0036
ID: cyst, Magnification : 2.5x



Sample: 0042_6 Host: 2015_0042
ID: cyst, Magnification : 6.3x



Sample: 318_2, Host: 2012_318
ID: cyst, Magnification : 12.0x